Genome composition and transposable element dynamics in Triticeae, the use of chloroplast sequences to date divergence times and transcriptome analysis of transposable element expression

Dissertation

zur Erlangung der naturwissenschaftlichen Docktorwürde

(Dr. sc nat.)

Vorgelegt der

Mathematisch-naturwissenschaftlichen Fakultät

der

Universität Zürich

von

Christopher Middleton

aus

United Kingdom

Promotionskomitee

PD Dr. Thomas Wicker (Leitung der Dissertation)

Prof. Dr. Beat Keller (Vorsitz)

Prof. Dr. Robert Dudler

Contents

1 Summary

The tribe of Triticeae contains some of the most important crop species, these include *Triticum aestivum* (wheat), *Hordeum vulgare* (barley) and *Secale cereale* (rye). Using next generation sequencing we want to explore phylogeny, divergence times and expression of transposable elements. Next generation sequencing (NGS) was conducted on twelve Triticeae species to produce an approximate genome coverage of 2% for each of the species analysed. Even with the comparatively low genome coverage it was still possible to gain several insights into different aspects of each genome. In the first part the transposable element composition of the different genomes was analysed, as these repetitive elements form approximately 80% of the genome. By far the largest contributing factor to the genomes was found to be the long terminal repeat (LTR) *Copia* retrotransposon *BARE1* which contributes between 10 and 13% of the total genomes. It was found that the abundance of different transposable element families varied strongly between taxa, with some elements only being highly abundant in some taxa, but not in others. We also wanted to assess the nucleotide diversity of the *BARE1* elements in Triticeae. It was found that nucleotide diversity is higher in the outbreeding taxa than the inbreeding taxa. Furthermore our data suggests that geographic isolation leads to lower nucleotide diversity of the *Angela* element.

Due to the high percentage of chloroplast reads in the samples, it was possible to assemble the chloroplast genome sequences of each of the taxa. This was used to produce a detailed phylogeny, estimate divergence times for each of the taxa and to identify *Aegilops speltoides* as the possible chloroplast genome donor of *Triticum aestivum*. Concise divergence times were estimated for each of the taxa, with *Hordeum vulgare* diverging from *T. aestivum* approximately 9 million years ago and the A, B and D genome donors of *T. aestivum* diverging within the last 3 million years.

Transcriptome data of wheat infected with *Blumeria graminis* or *Mycosphaerella graminicola* was investigated and revealed differences in expression patterns of a number of wheat transposable elements. Several elements were investigated at various time points during infection and this gave evidence that different transposable elements show different expression

patterns during the course of infection. Our data indicate that TEs show a wide variety of expression patterns, but also evidence for them being silenced by the host.

Zusammenfassung

Die Gruppe der Triticeae umfasst einige der wichtigsten Getreidespezies der Welt, darunter *Triticum aestivum* (Weizen), Hordeum vulgare (Gerste) and Secale cereale (Roggen). Mit Hilfe von Next-Generation-Sequencing (NGS) wollten wir die Stammesgeschichte, die Zeiten der evolutionaren Aufspaltung der spezies und die Expression von transposablen Elementen (TEs) untersuchen. NGS wurde an zwolf Tritceae-Species durchgefuhrt, was eine Abdeckung des Genoms von ca. 2% ermoglichte. Trotz mit dieser relativ geringen Abdeckung war es moglich wichtige Einblicke in verschiede Aspekte des Genoms zu gewinnen. Im ersten Teil des Projekts wurde die TE-Zusammensetzung der verschiedenen Genome analysiert, da diese Elemente etwa 80% des Genoms ausmachen. Mit Abstand am starksten vertreten waren Long-Termin-Repeat (LTR) Retrotransposons der *BARE1*-Familie, welche zwischen 10% und 13% der Genome ausmachten. Wir entdeckten, dass die Beitrage der verschiedenen TE-Familien zum Gesamtgenom sich zwischen den Spezies stark unterschieden. Einge TE-Familien waren stark vertreten in der einen Spezies wahrend sie in einer anderen praktisch vollig abwesend sein konnten. Wir wollten auch die Sequenzdiversitat der *BARE1*-Elemente im Genom quantifizieren und fanden, dass die Diversitat geringer ist in Spezies die selbstbefruchtend sind. Ausserdem fanden wir Hinweise darauf, dass geografische Isolation zu geringerer Sequenzdiversitat fhhren kann. Da die sequenzierten DNA-Extrakte viel Chloroplasten-DNA enthierten, konnten wir Chloroplasten-Genome fur alle untersuchten Spezies zusammen setzen. Diese wurden dann verwendet, um eine genaue Stammesgeschichte abzuleiten und um die Zeiten der evolutionaren Aufspaltung der einzelnen Triceae-Spezies zu berechnen. Wir konnten auch klar Zeigen, dass der Donor des B-genoms von *T. aestivum* ein naher Verwandter von *Aegilops speltoides* sein muss. Wir schatzten, dass sich H. vulgare vor ca. 9 Millionen Jahren von der Weizen/Roggen Linie abgespalten hat, warend die Donoren der A-, B- und D-Genome von Weizen sich vor ca. 3 Millionen Jahren von einander abspalteten. Transkriptomdaten von Weizen, der mit *Blumeria graminis* und *Mycosphaererella graminicola* infiziert worden war brachte eine Vielfalt an Expressionsmustern in TEs and den Tag. Hier fanden wir Hinweise darauf, dass einge TE-Familien komplexe Expressionsmuster zeigen,

andere aber vom Wirt-Genom in ihrer Expression unterdruckt werden.

2 General introduction

The Triticeae are a diverse group of grasses within the subfamily pooideae and include several major crop species such as *Triticum aestivum* (bread wheat), *Hordeum vulgare* (barley) and *Secale cereale* (rye). Several minor crop species are also found within this tribe, including *Triticum spelta* (spelt wheat), *Triticum monococcum* (einkorn wheat), *Triticum turgidum ssp. durum* (durum wheat) and Triticale, a hybrid of wheat and rye. Bread wheat is one of the most important crop species worldwide, providing staple nutrition for approximately 30% of the global population. Bread wheat production reached 701 million tons in 2011 (FAO, 2011), making it the third largest crop in terms of production after maize (851 million tons) and rice second with a global production of 722 million tons. Barley is a lesser crop in comparison and 123 million tons were produced in 2010 (FAO, 2012).

The roots of the domestication of the Triticeae can be traced to the fertile crescent, a region in the middle East that includes Iran, Iraq, Kuwait, Turkey, Syria, Jordan and Israel. It was in this region that it is presumed that the initial domestication of many cereal crops began (Kihara, 1944; Feldman *et al.*, 1995; Devos *et al.*, 2005; Kilian *et al.*, 2007a; Bordbar *et al.*, 2011). It was during the early Neolithic period approximately 8,000 - 15,000 years ago, when populations of people began to make the switch from the former hunter gatherer to a more structured society that coincided with the earliest crops being domesticated (Eckardt, 2010). The earliest Triticeae species to be domesticated in this region were *H. vulgare* and *T. monococcum*. Probably, *T. monococcum* was the earliest to be domesticated and it was domesticated from its wild progenitor *T. boeoticum*. This is believed to have occurred approximately 12,500 years ago (Kilian *et al.*, 2007a). Archaeological evidence shows a shift in the seed size from the smaller seeds in the wild progenitor *T. boeoticum* to the the larger seed size seen in the domesticated *T. monococcum*. However, the diploid *T. monococcum* was later abandoned as a crop with the introduction of *T. aestivum* approximately 8,000 years ago (Kilian *et al.*, 2007a).

T. aestivum is an allohexaploid, with the genome denotion AABBDD and was formed from

the hybridisation of three distinct genomes. The first hybridisation event occurred between the diploid species *Triticum urartu* (AA) and possibly *Aegilops speltoides* or (SS) to form the tetraploid *Triticum turgidum* ssp. *dicoccon* (AABB) (emmer wheat), (Mori *et al.*, 1995; Huang *et al.*, 2002b; Dvorak *et al.*, 2005). Although, the exact nature of the origin of the B genome donor is unknown it is largely speculated to be a now extinct ancestor of *Ae. speltoides* (Kimber, 1974; Petersen *et al.*, 2006; Kilian *et al.*, 2007b). A further hybridisation event occurred between the tetraploid *T. turgidum* and the goat grass *Aegilops tauschii* (D genome) approximately 8,000 years ago. This event most likely happened in the early cultivated tetraploid *T. turgidum ssp. dicoccon* (emmer) fields to produce the hexaploid *T. aestivum* (Akhunov *et al.*, 2003; Edwards *et al.*, 2010; Tomita *et al.*, 2010).

It is thought that the domestication of *T. monococcum* and *T. turgidum* ssp. *dicoccon* occurred at different locations and points in time (Feldman *et al.*, 1995). Grain size was not the only important trait selected for during domestication. Non-brittle rachis was also a desired trait when compared to the brittle form in the wild species and resulted in the ears retaining the seeds upon ripening, producing a more free threshing variety (Salamini *et al.*, 2002; Özkan *et al.*, 2005)

The movement of people further North towards the end of the ice age allowed new regions to be colonised and enabled these grasses to be domesticated throughout the middle East and Europe. A gradual process of improvement began and traits were selected based on the response to environmental factors, such as day length and the time taken for the grain to mature. This was however a slow process of gradual improvement over the past several thousand years and it has only been in the last hundred years or so, with a greater understanding of genetics and a better grasp of the nature of heredity that has led to a vast improvement of crops. The so called green revolution began in the 1940s and continued until the 1970s. The green revolution was instigated by improvements in not only the selection for more desirable traits in the plant itself, but also in better land management including the introduction of fertilisers and pesticides and this rapidly led to large increases in yield (Peng *et al.*, 1999).

2.1 Genome size and variation within plants

The genome sizes of the Triticeae are large when compared to other plant species where the genome sequence is already available, with the genome size of the Triticeae typically being between 3,500 - 8,500 mega base pairs (Mbp). For example the angiosperm model plant species *Arabidopsis thaliana*, which was the first completely sequenced plant genome, has a genome size of approximately 120 Mbp and a chromosome number of n = 5 (AGI, 2000). The variation in genome size of dicotyledonous plants is much smaller than that of the monocotyledonous plants (Chaw *et al.*, 2004). The size of dicot genomes generally varies from between approximately 120 Mb in *A. thaliana* to approximately 975 Mb in *Glycine max*. A much larger range in genome sizes is found in the monocots were the genome sizes can vary from between a few hundred megabases to several thousand megabases.

The release of the *A. thaliana* genome was closely followed two years later by the complete sequences of the first monocotyledon grass species *Oryza sativa* L ssp. *japonica* (Yu *et al.*, 2002) and *Oryza sativa* L ssp. *indica* (Wang *et al.*, 2002). The rice genome is one of the smallest of the crop species. At approximately 430 Mbp it is three and a half times larger than that of *A. thaliana*. The sequencing of the rice genomes was followed by the release in 2010 of the first Pooideae subfamiliy member *Brachypodium distachyon*, this is a wild annual grass found in the Middle East and Mediterranean, with a genome size of approximately 272 Mbp (Initiative, 2010). The largest plant genome completely sequenced so far is that of maize with a genome size of 2,300 Mbp Schnable *et al.* (2009).

2.2 Genome sequencing in plants

The method used to sequence the genome of *Oryza sativa* L ssp. *indica* differed to that of the *Arabidopsis* genome in that this rice species was sequenced using a whole-genome shotgun (WGS) approach. In this approach the genome is randomly broken up into smaller segments prior to being sequenced. Where as, the *Arabidopsis* genome and the genome of *Oryza sativa* L ssp. *japonica* was sequenced by constructing a bacterial artificial chromosome (BAC) library

of the whole genome, then these BAC clones where sequenced by shotgun sequencing. However, with the introduction of next generation sequencing, the cost and time taken to produce complete genome sequences has decreased (Shendure *et al.*, 2008).

The highly repetitive nature of the Triticeae genomes has made producing a complete genome sequence elusive. The repetitive fraction of the Triticeae species is approximately 80% (Bennett *et al.*, 1976, 2011). However, draft sequences have become available for two of the *T. aestivum* genome donors, *T. urartu* A genome (Ling *et al.*, 2013) and *Ae. tauschii* D genome (Jia *et al.*, 2013). Furthermore, there are currently projects in the pipeline to produce a completely assembled genome sequence of *H. vulgare* (International barley sequencing consortium), with a partial sequence covering mainly the gene space being published recently (Consortium *et al.*, 2012) and the international wheat genome sequencing project (IWGSC), has plans to completely sequence the the large 16 Gbp genome of *T. aestivum*. These projects were only able to come to fruition due to the introduction of next generation sequencing (NGS) strategies. Since 1982 the amount of genetic sequence available in genbank has doubled approximately every 18 months. However, since the inception of NGS the cost and time taken to produce high quality datasets has massively been reduced in recent years. Several new methods were developed, these include Illumina, 454, Solid and more recently PacBio and ion torrent. These have enabled researchers to establish protocols for the sequencing assembly of large complex genomes at very low costs (Shendure *et al.*, 2008).

2.3 Transposable elements

Transposable elements constitute a major portion of the genomes of many higher eukaryotes, in comparison with protein coding sequences, which typically only make up a much smaller proportion of the genome. For example in humans the protein coding sequences make up approximately 3 % of the genome, whereas 50% of the genome is made up from transposable elements (Cavalli-Sforza, 2005). As TEs make up such a large proportion of the Triticeae genomes, it is important they are discussed in more detail. TEs are found in almost all organ-

isms and a unified classification system had to be established to account for the large amount of variation seen between different classes of TEs (Wicker *et al*,. 2007). TEs were initially divided into two classes based on the mechanism of transposition. The two classes were further subdivided into 9 orders and 29 superfamilies, with class I TEs containing 5 of these orders and 17 superfamilies and class II TEs making up the rest of the classification system and are made up of 4 orders and 12 superfamilies. By classifying TEs into this system it has enabled researches to identify the type of TE more rapidly and upon discovery of a new element it can be assigned into the correct class. This method has also resulted in the formation of a large database of known Triticeae TEs and this enables rapid identification of TEs from large genomic datasets (TREP) (http://wheat.pw.usda.gov/ITMI/Repeats/).

Transposable elements have been refereed to as "junk DNA" as it was previously believed that these intergenic sequences were not involved in the function of maintaining the genome and were more akin to hitchhikers hiding in the genome to escape degradation. This outdated terminology is in stark contrast to what is now known about transposable elements and their role in many genomic processes. TEs have an influence in genome dynamics, including genome size and by influencing the expression of genes by either direct insertion into the gene or through insertion into the promoter sequences (Vicient *et al.*, 1999; Vitte *et al.*, 2005; Giovanni *et al.*, 2008). The adverse effect that TEs can have upon gene expression is tightly controlled and the cell has two distinct mechanisms to reduce the impact that TEs can have on the genome. One method is through the methylation of cysteine residues within the TE sequence or through modification of histones leading to epigenetic silencing (Jaenisch *et al.*, 2003). Chromatin remodeling through histone modification results in the silencing of TEs, it was found in *Arabidopsis* that a mutant for the protein DDM1, which is involved in chromatin modification results in a burst of TE activity (Tsukahara *et al.*, 2009). The second way in which TEs are silenced is through the RNAi silencing pathway, resulting in the formation of small RNAs that complement the transcribed TE sequence and target the RNA for degradation (Baulcombe, 2004; Qi *et al.*, 2006).

A large number of TE families is found in the genomes of Triticeae, this could possibly be

the result of minor changes in the sequence of a particular TE, resulting in the formation of a new TE family that is unrecognised by the silencing machinery and can rapidly colonise new regions of the genome, before the mechanisms which silence them catch up.

It has also been found that TEs proliferate during times when the organism is under stress. For example, during drought stress or the invasion by a pathogen, these can lead to the proliferation of TEs within the genome. For example in the wild barley *H. spontaneum* changes in climate can result in changes in the abundance of the *Copia* element *BARE1* (Kalendar *et al.*, 2000). These mechanisms often result in changes in the TE variation found within a genome with some families becoming more prolific and others decreasing in abundance, with these differences being seen between closely related species (Kalendar *et al.*, 2000; Vitte *et al.*, 2005; Piegu *et al.*, 2006).

2.4 Class I transposable elements

Class I TEs are the major colonisers of Triticeae genomes and they transpose via a RNA intermediate, where they are first copied to RNA, before being reverse transcribed into DNA and reinserted in the genome in a different location. These retrotransposons form a large proportion of eukaryotic genomes and can be divided into two groups long terminal repeat (LTR) and non-LTR retrotransposons. LTR retrotransposons are divided into two groups the *Copia* and *Gypsy*. Non-LTR retrotransposons are classified into long interspersed nuclear elements (LINES) and short interspersed nuclear elements (SINES). Retrotransposons are similar in their structure and life cycle to retroviruses, which are possibly derived from these, with the only difference being the *ENV* gene that encodes for the envelope protein, to enable virus mobility.

Upon insertion into a new location in the genome a 4-6 bp target site of duplication (TSD) is produced. In LTR retrotransposons the long terminal repeat at either end, flanking the coding region of the element vary in length from between approximately 100 bp to several thousand bp, with the sequence of the LTR starting with 5'-TG 3' and ending with 5'-CA 3'. Both the

Copia and *Gypsy* elements encode a number of genes that are involved in transposition of the element in the host. These include two open reading frames (ORFs). The first contains *gag* which encodes the structural protein forming the nucleocapsid and the second is *pol* which encodes the enzymatic functions related to this polypeptide. The pol polypeptide contains a reverse transcriptase (RT), a RNAseH and an intergrase (INT). RT carries out reverse transcription of the newly formed RNA and the DDE intergrase INT, which inserts the newly formed LTR reterotransposon copy into the genome. The difference between the *Gypsy* and *Copia* superfamilies is largely down to the order of these genes, with the integrase occurring at the end of the ORF in *Gypsy* elements and before the reverse transcriptase gene in *Copia* elements. LTR retrotransposons can vary considerably in size from a few hundred bp to approximately 25 kb in the case of the *Ogre* element. In Triticeae LTR reterotransposons are the largest contributor of TEs in the genome and it was found in *H. vulgare* that they form approximately 50% of the genome (Wicker *et al.*, 2009b).

The non LTR retroelement LINE encodes two ORFs,l with ORF1 being composed of an RNA binding protein and ORF2 encoding genes involved in endonuclease activity and reverse transcription. Transcription of LINEs usually begins at a promoter located in the 5' end of the element, whereas SINEs contain an internal RNA *poll III* promoter located near the 5' end of the element. LINEs are typically several kb in length, SINEs however are much smaller and range in length from 90 - 500 bp. In both LINEs and SINEs transcription is terminated by a poly(A) sequence repeat. In Triticeae LINEs and SINEs make up a much lower proportion of the genome compared to LTR retrotransposons, with generally less than 10% contributing to the whole genome of Triticeae (Wicker *et al.*, 2009b; Senerchia *et al.*, 2013).

2.5 Class II transposable elements

Class II transposons are found in all eukaryotes and are DNA transposons that are directly excised from their position in the genome before inserting into another location elsewhere. Class II transposons exist as two subclasses. Subclass I TEs are generally characterised by

the presence of a terminal inverted repeat (TIRs) at either end of the element, these are found in all 9 subfamilies of this element. The nine subfamilies are defined by the length of the TIR and the TSD. The transposition of these elements is mediated through a transposase enzyme that recognises the TIR, cutting both strands. Transposons in subclass I include Tc1/*mariner*, hAT, Mutator, Merlin, *piggyback* and CACTA.

MITEs (miniature inverted repeat transposons) also make up a relatively large proportion of plant genomes. These elements are derived from class II non autonomous elements, they have a typical size of between 100 and 600 bp and contain terminal inverted repeats. MITEs are generally located in or near genes (Bureau *et al.*, 1994) and like other nonautonomous elements they are dependent upon the transposases encoded by autonomous elements for transposition. Most of the large number of MITEs present in plant genomes can be divided into two groups, these are *Tourist*-like MITEs and *Stowaway*-like MITEs, with *Tourist*-like MITES, being derived from *Harbingers* and *Stowaway*-like MITES, being derived from *Mariners* (Bureau *et al.*, 1994; Yang *et al.*, 2009). These two groups make up the largest numbers of MITEs in the genomes of several species. In rice the *Stowaway* MITEs makes up around 2% of the genome with the *Tourist* MITEs providing approximately 6% of the rice genome.

As class II elements are directly cut and moved to another location in the genome, their increase in opulence within the genome usually comes about due to transposition during chromosome replication, where they are excised from an already replicated chromosomal location to a region that the replication fork has not yet passed.

Subclass II elements in this class of transposable elements transpose via a different mechanism that does not require direct cutting of both strands at either end of the TIR sequence, instead they transpose through the displacement of one strand or in the case of *Helitrons* through a rolling circle mechanism (Kapitonov *et al.*, 2007). *Helitrons* can be distinguished by its enzymatic functions and contains domains for rolling-circle replication initiator (Rep) and a DNA helicase (Hel). The Rep domain is approximately 100 amino acids long and is involved in endonucleolytic cleavage, DNA transfer and ligation. The Hel domain produces a DNA helicase of approximately 400 amino acids and belongs to superfamily 1 (Kapitonov *et al.*, 2001).

2.6 Evolutionary studies in Triticeae

Many phylogenetic studies have been conducted to try and understand the relationships between the many species of the Triticeae. These earlier investigations focused on using a subset of genes, rDNA or smaller intergenic sequences to determine phylogeny (Gaut, 2002). Since the introduction of next generation sequencing new methods were developed to utilise larger genome fractions to not only determine phylogeny, but also to date the divergence of the individual species.

Several statistical methods were introduced to draw phylogenetic trees, these are usually based upon sequence alignment or through the calculation of pairwise distances. However, these distance based methods involving pairwise analysis have largely been replaced as computational power has increased and has lead to the use of aligned sequences as an alternative. Maximum parsimony (MP) is one statistical method to draw phylogenetic trees from aligned sequences. MP draws the tree using the smallest number of substitutions along the branches, that is it looks at each individual sequence and compares the nucleotide substitutions between the sequences and then draws the tree based on the simplest explanation that fits the evidence. However, MP does have some draw backs when compared to other methods. This becomes more apparent when larger datasets are used and a more heuristic approach is taken to find the best fitting tree. Another problem with using this approach is that MP links taxa that have long branches, this is due to the branches that share many substitutions, but does not necessarily share the same common ancestor (Rutschmann, 2006; Duchene *et al.*, 2013a).

As the computational power has increased better methods were developed to ascertain phylogeny. Maximum likelihood is one of these methods, it relies quite heavily on complex statistics and hence the increased need in computational ability. The maximum likelihood method calculates the sums of all nucleotide states in the tree for the positions in the alignment, that is it measures all nucleotide substitutions and selects the tree with the highest likelihood as

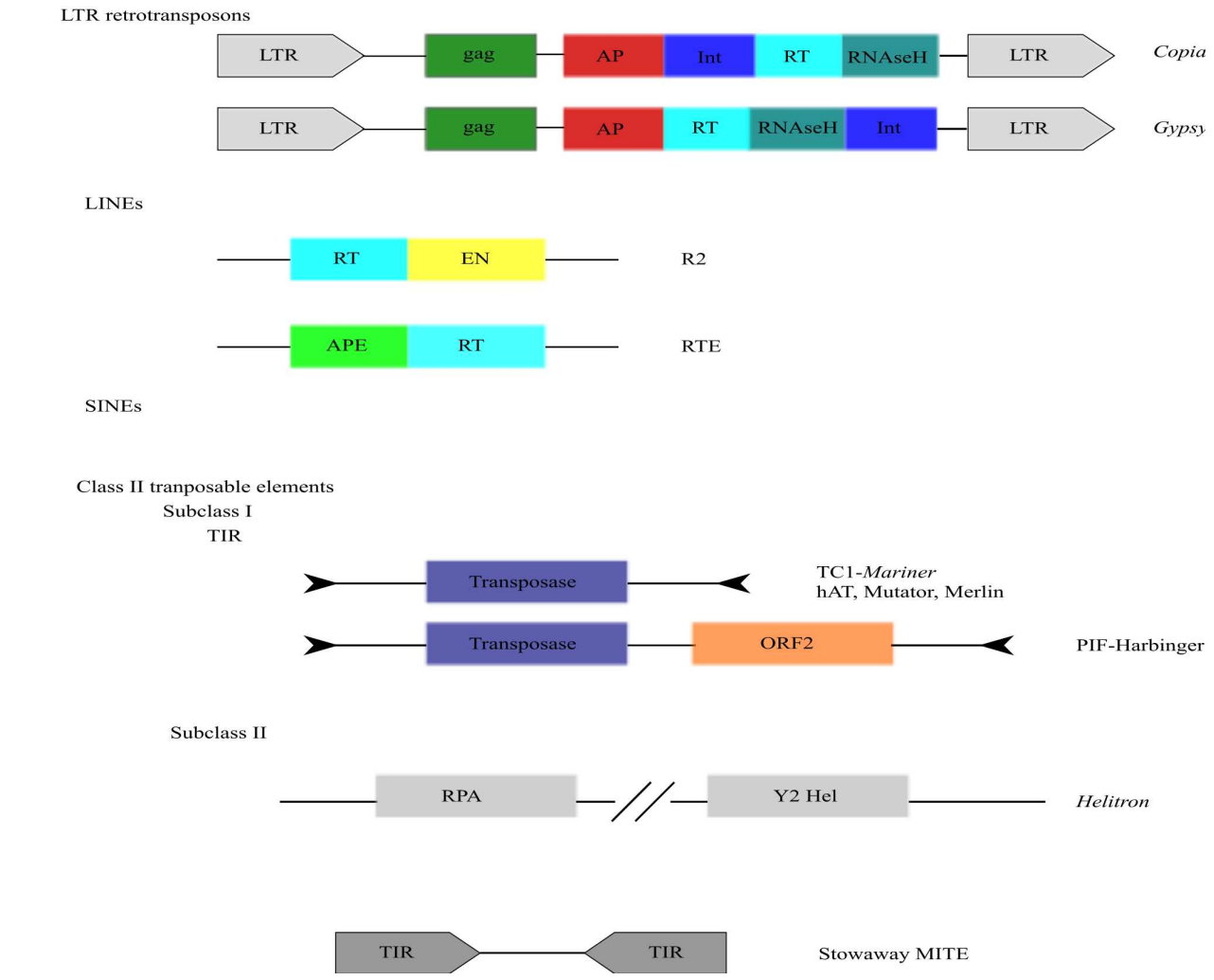


Figure 1. Diagram showing the structure of each of the most abundant transposable element superfamilies in plants.

the best candidate. The maximum likelihood method requires the use of a substitution model to take into account the different permutations involved in DNA base substitutions (Mar *et al.*, 2005; Rutschmann, 2006). The most commonly used and complex of these is the general time reversible model. This model uses all possible variations in mutations to calculate the best substitution rate for each individual base in the alignment.

An alternative method to maximum likelihood is to use Bayesian inference, this method has gained popularity in recent years and becoming more common in its use for evolutionary studies (Mar *et al.*, 2005; Rutschmann, 2006). In Bayesian inference the posterior probability of the tree being correct is associated with the prior probability, which takes into account all of the information given, such as the evolutionary model and data (Mar *et al.*, 2005).

2.7 Estimating divergence times

Dating the divergence of species is based upon substitution rates of nucleotide sequences and there are several methods to calculate the substitution rates. A large number of models were developed to account for the different types of mutations. Bootstrapping is fundamentally important in validating the correctness of the tree. This tests the probability that the branches lie in the correct position on the tree and normally one thousand repetitions are conducted to determine the correctness of the tree. From this information on the branch lengths and substitution rates it is possible to date the divergence of each of the species in the tree. This is usually done based on a calibration date, with this typically coming from a predetermined date and is normally based upon a divergence time of an earlier species identified from the fossil record. All of this information combined can be used to determine the molecular dates of divergence (Mar *et al.*, 2005; Rutschmann, 2006).

The region of the genome used in dating the divergence times is also important for obtaining accurate times of divergence. For example care has to be taken when using gene sequences in calculations as these are under a strong selection pressure. It has become standard practice when using gene sequences to date the divergence to use the synonymous mutation rate

to obtain the date. Synonymous substitutions typically occur in the third position of the codon and don't affect the amino acid sequence and are therefore less likely to be under a selection pressure, when compared to nonsynonymous substitutions which generally occur in the first and second positions of a codon. However, Buchmann *et al* 2012 showed that synonomous sites in grass genes are not completely free from selection pressure and this can lead to over estimation of divergence times. The general consensus when determining phylogeny and divergence times, has been to use either genes, intergenic sequences or highly repetitive sequences such as rDNA (Huang *et al.*, 2002a; Akhunov *et al.*, 2007; Chalupska *et al.*, 2008). However, more recently with the advancement of high throughput sequencing it is becoming more common to use larger sequences, such as those from chloroplasts and mitochondria (Nikiforova *et al.*, 2013).

2.8 Aims of this study

The main goals of this study were to utilise 454 sequencing to study the relationships between different members of the Triticeae tribe and to analyse the transposable element composition in the genomes of different Triticeae species. As TEs form such a large part of the genome, we wanted to assess if there are differences in the families that make up a large part of each of the genomes. Another goal of this study was to draw a detailed phylogeny of the Triticeae species presented here and to accurately date the times of divergence for each of the species used. A further aim was to use transcriptome data from *T. aestivum* infected with a pathogen to try and gain a better understanding of TE dynamics under biotic stress.

3 Comparative analysis of genome composition in Triticeae reveals strong variation in transposable element dynamics and nucleotide diversity

Christopher P Middleton, Nils Stein, Beat Keller, Benjamin Kilian and Thomas Wicker[1]

[1]Corresponding author

Published in Plant Journal (2013), Volume 73. Issue 2. Pages 347-356

3.1 Summary

A 454 sequencing snapshot was utilised to investigate the genome composition and nucleotide diversity of transposable elements (TEs) for several Triticeae taxa, including *Triticum aestivum*, *Hordeum vulgare*, *Hordeum spontaneum* and *Secale cereale* together with relatives of the A, B and D genome donors of wheat, *Triticum urartu* (A), *Aegilops speltoides* (S) and *Aegilops tauschii* (D). Additional taxa containing the A genome, *Triticum monococcum* and its wild relative *Triticum boeoticum*, were also included. The main focus of the analysis was on the genomic composition of TEs as these make up at least 80% of the overall genome content. Although more than 200 TE families were identified in each species, approximately 50% of the overall genome comprised 12–15 TE families. The *BARE1* element was the largest contributor to all genomes, contributing more than 10% to the overall genome. We also found that several TE families differ strongly in their abundance between species, indicating that TE families can thrive extremely successfully in one species while going virtually extinct in another. Additionally, the nucleotide diversity of *BARE1* populations within individual genomes was measured. Interestingly, the nucleotide diversity in the domesticated barley *H. vulgare* cv. Barke was found to be twice as high as in its wild progenitor *H. spontaneum*, suggesting that the domesticated barley gained nucleotide diversity from the addition of different genotypes during the domestication and breeding process. In the rye/wheat lineage, sequence diversity of *BARE1* elements was generally higher, suggesting that factors such as geographical distribution and mating systems might play a role in intragenomic TE diversity.

3.2 Introduction

Hexaploid wheat (*Triticum aestivum*) is a major crop world-wide. It is a member of the Triticeae tribe, which also includes other economically important species such as *Hordeum vulgare* (barley) and *Secale cereale* (rye). The barley lineage also includes wild taxa such as *Hordeum vulgare* ssp. spontaneum. The wheat lineage includes the hexaploid *T. aestivum* as well as its diploid genome donors *Triticum urartu*, *Aegilops speltoides* and *Aegilops tauschii*. In addition it includes wild einkorn wheat (*Triticum monococcum ssp. boeoticum*) and its domesticated descendant *Triticum monococcum ssp. monococcum*, which are closely related to *T. urartu*. Triticeae species probably originated in the Fertile Crescent, which includes Iran, Iraq, south-east Turkey, Syria, Lebanon, Jordan and Israel (Kihara, 1944; Feldman *et al.*, 1995; Devos *et al.*, 2005; Kilian *et al.*, 2007b; Bordbar *et al.*, 2011). However, the divergence times and phylogenetic relationships, especially between bread wheat and closely related taxa, is not fully understood. *Hordeum vulgare* and the wheat/rye lineage were predicted to have diverged 10–15 million years ago (Ma), while wheat and rye diverged approximately 5–11 Ma (Chalupska et al., 2008). *Triticum urartu*, *Ae. speltoides* and *Ae. tauschii* were predicted to have diverged from each other between 2 and 6 Ma (Huang *et al.*, 2002a; Akhunov *et al.*, 2003; Chalupska *et al.*, 2008).

Triticum aestivum has a total hexaploid genome size of approximately 16–17 Gb (Rees *et al.*, 1965; Bennett *et al.*, 1976). Wheat is an allohexaploid and formed of three genomes, denoted A, B and D. The haploid sizes of these genomes in *T. aestivum* are very similar to those of the other Triticeae members, which are generally 3500–8500 Mb (Eilam *et al.*, 2007; Özkan *et al.*, 2010; Bennett *et al.*, 2011). The complete genomic complement of T. aestivum AABBDD was formed from the hybridisation of three diploid ancestors (Akhunov *et al.*, 2003; Edwards *et al.*, 2010; Tomita *et al.*, 2010) The first hybridisation event is estimated to have occurred between 0.20 and 1.3 Ma (Mori *et al.*, 1995; Huang *et al.*, 2002a; Dvorak *et al.*, 2005), between *T. urartu* (AA) and possibly *Ae. speltoides* (SS), to form the tetraploid species *Triticum turgidum ssp. dicoccoides* (AABB), wild emmer wheat (Dvorak and Zhang, 1990; Dvorak *et al*., 1993; Akhunov *et al*., 2005; Kilian *et al*., 2007b). The genome of domesticated emmer wheat *T.*

turgidum ssp. dicoccon was further complemented by the addition of the D genome from *Ae. tauschii* approximately 8000 years ago to form the hexaploid *T. aestivum* (Kihara, 1944; McFadden *et al.*, 1946; Feldman *et al.*, 1995; Devos *et al.*, 2005; Dubcovsky *et al.*, 2007; Kilian *et al.*, 2007b; Bordbar *et al.*, 2011).

The large genome size of the Triticeae members and the presence of high numbers of repetitive elements which make up at least 80% of the whole genome complement (Bennett *et al.*, 1976; Hollister *et al.*, 2007), have made sequencing the genomes of these species extremely challenging. As transposable elements (TEs) form such a large component of the genome, previous studies have generally focused on the contribution of different TE families to single species (Wicker *et al.*, 2009b; Rebollo *et al.*, 2010; Tenaillon *et al.*, 2011), while the differences in TE families between species are less well understood. Transposable elements come in two classes according to their method of transposition: Class I, which copy themselves via a RNA intermediate, before the newly synthesised element is inserted into a different region of the genome, and Class II, which transpose in a copy–paste mechanism, as they are directly cut from their position in the genome and reinserted elsewhere (Feschotte *et al.*, 2002; Casacuberta *et al.*, 2003). There are large variations in the amount and the copy number of each element, when comparing different genomes (Kidwell, 2002). For example the genome of Arabidopsis thaliana contains approximately 10% TEs, whereas in most grass species TEs comprise between 50% (rice) and 80% (wheat) of the entire genome complement (Feschotte *et al.*, 2002). Transposable elements play a role in a number of evolutionary processes, including insertion into protein-coding genes, illegitimate recombination and chromosome breakage (Slotkin *et al.*, 2007). Anyone of these can have an influence over the fitness of the host. There are several mechanisms to control the level of transposition, including post-transcriptional silencing and methylation, However, these systems can be exacerbated during times of abiotic stress, leading to a proliferation of elements in the genome (Vicient *et al.*, 1999; Todorovska, 2007).

Nucleotide diversity, π, represents the average sequence divergence between all homol-

ogous sequences among all individuals in a given set for comparison (Nei *et al.*, 1979). It is often used to infer the presence of past population bottlenecks in studies of domestication genetics, because when a population goes through a bottleneck, the allelic diversity in the population is diminished, and π is thus expected to be small. A rare case of a cereal in which there have been no recent breeding bottlenecks was described by (Kilian *et al.*, 2007a), where no reduction of nucleotide diversity at all was found. The absence of a domestication bottleneck is in contrast to the conclusions of studies of domestication in intensively bred crop species, where claims for domestication bottlenecks are commonplace (Buckler *et al.*, 2001; Doebley *et al.*, 2006; Kilian *et al.*, 2006). In that study, Kilian *et al*. (2007a) investigated nucleotide variation at 18 loci from 92 domesticated einkorn lines compared with 321 lines from wild populations. Several insights into domestication history emerged from that study. One of the most important insights was that wild einkorn is not really a single homogeneous population, rather it underwent a natural process of genetic differentiation prior to domestication, resulting in three distinct wild einkorn races. These three races, which were designated as α, β and γ, are genetically distinct both at the level of their haplotypes across 18 loci studied and at the level of their amplified fragment length polymorphism fingerprints. One of those races, wild race β, is genetically much more similar to domesticated einkorn, hence it is the race, or genotype, that was exploited by humans during domestication. Race β occurs only in the Karacadag and Kartal-Karadag Mountains in south-east Turkey today. A second major surprise in the findings of Kilian *et al*. (2007a) was that nucleotide and haplotype diversity in domesticated einkorn was found to be higher than in the β race. However, very little is known about the nucleotide diversity within transposable families within a genome.

Mating systems may also have an influence on the numbers of TEs and the nucleotide diversity of the elements within a genome. Beside the outbreeder *S. cereale*, two predominantly oubreeding species are known in the Triticum–Aegilops group within the Triticeae tribe: *Ae. speltoides* and *Aegilops mutica* (Kimber, 1987; Kilian *et al.*, 2007b, 2011). All other taxa including, for example, *T. urartu*, einkorn wheat and *Ae. tauschii* are inbreeders. Mating sys-

tems have been found to influence molecular evolution, reducing the levels of polymorphism and affecting the effective population size (Haudry *et al.*, 2008). It has been stipulated that inbreeding can reduce the effective population size by as much as 50% (Pollak, 1987), therefore it can be extrapolated that there would also be a reduction seen in nucleotide diversity dependent upon mating system, with lower diversity seen in inbreeders compared with outbreeders. Transposable element families also represent populations inside genomes with subfamilies and hardly ever are two copies of a family absolutely identical. Little is known about the nucleotide diversity of TE families within a genome as there has been no quantitative survey. The TE diversity within a genome has not been studied before. Therefore it is not known what influence domestication, mating system or geographical isolation will have on the TE diversity within a genome.

Next-generation sequencing provides new opportunities due to the large volume of datasets it can produce. Many studies have been conducted using 454 sample sequencing as a platform (Macas *et al.*, 2007; Swaminathan *et al.*, 2007; Mardis, 2008); Wicker et al. (2009) utilised this method to analyse the TE composition of the barley genome. The main finding was that a small number of TE families contribute to more than 50% of the genome, with the vast majority of these pertaining to the class I long terminal repeat (LTR) retrotransposons. It has been noted before that the *BARE1* clade in barley and the *Angela/Wis* clade in wheat form approximately 10% of the genome (Vicient *et al.*, 1999; Kalendar *et al.*, 2000; Soleimani *et al.*, 2006; Wicker *et al.*, 2009b). The study conducted by Wicker *et al*. (2009) looked at the barley repetitive fraction, with a comparison being made with a limited dataset from *T. aestivum*, suggesting that TE compositions of Triticeae genomes vary between taxa, for example the *Gypsy* element *BAGY2* was more abundant in the barley than in the wheat taxa (Wicker et al., 2009). However, no broad survey of an entire tribe has been conducted yet.

Here we used 454 sequencing to obtain between 3 and 5% genome coverage for nine Triticeae species including the A, B and D genome donors of *T. aestivum*. We wanted to ad-

dress the following questions: (i) Are there differences in the composition of the genome, in particular regarding the abundance and variation of transposable elements between the taxa? (ii) Are there differences in the level of nucleotide diversity of particular transposable elements within the genome, with attention being paid to different factors such as the domestication process, geographical distribution and mating type.

3.3 Results

454 sample sequencing and characterisation of genome compositions

454 titanium 7 kb paired end sequencing was done to produce a genome sequence coverage of between 2 and 5% for each of *S. cereale*, *T. urartu*, *Ae. speltoides* and *Ae. tauschii*. The datasets resulted in approximately 441,000–546,000 reads, with an average size of between 260 and 300 bp. In addition, 454 sequences of the *Triticum* taxa *T. monococcum ssp. boeoticum* (hereafter *T. boeoticum*) and *T. monococcum ssp. monococcum* (hereafter *T. monococcum*) and the *Hordeum* taxa *H. vulgare* cv. Barke and *H. vulgare ssp. spontaneum* (hereafter *H. spontaneum*) accessions FT11 and FT462 were analysed. The sample of these additional taxa consist of approximately 450 000–1 300 000 reads with average read lengths of 295–396 bp. The three taxa *T. urartu*, *T. boeoticum* and *T. monococcum* all contain the A genome and will be referred to in the text as the 'A genome taxa'. Furthermore, 500 000 publicly available *T. aestivum* cv. Chinese Spring 454 sequences with an average size of 415 bp were also included in the analysis (http://www.cerealsdb.uk.net/CerealsDB/Douments/) (Table 1). The reads for each of the species were classified using BLAST searches against different databases, which included Triticeae repeats TREP10 (known TEs) and a BLASTX search against PTREP11 (TE protein sequences) as well as organelles and genes. Transposable elements made up the largest pro- portion of reads for each of the grass species, with between 64.2 and 71.2%. The *T. urartu* sample contained the highest number of identified TEs with a total of 72.3% of all reads (Figure 2).

A small number of TE families make up a large proportion of Triticeae genomes

Further investigation of the TE reads was conducted in order to identify the proportions of different TE families present in each of the genomes. All taxa analysed were found to contain between 226 and 241 different TE families. A large number of the identified TE families are present in very low copy numbers in the genomes, and approximately 15 different families make up at least 50% of the genome complement for each species (Figure 3). Approximately 70% of the characterised TE reads belong to the *Gypsy* superfamily of retrotransposons. In

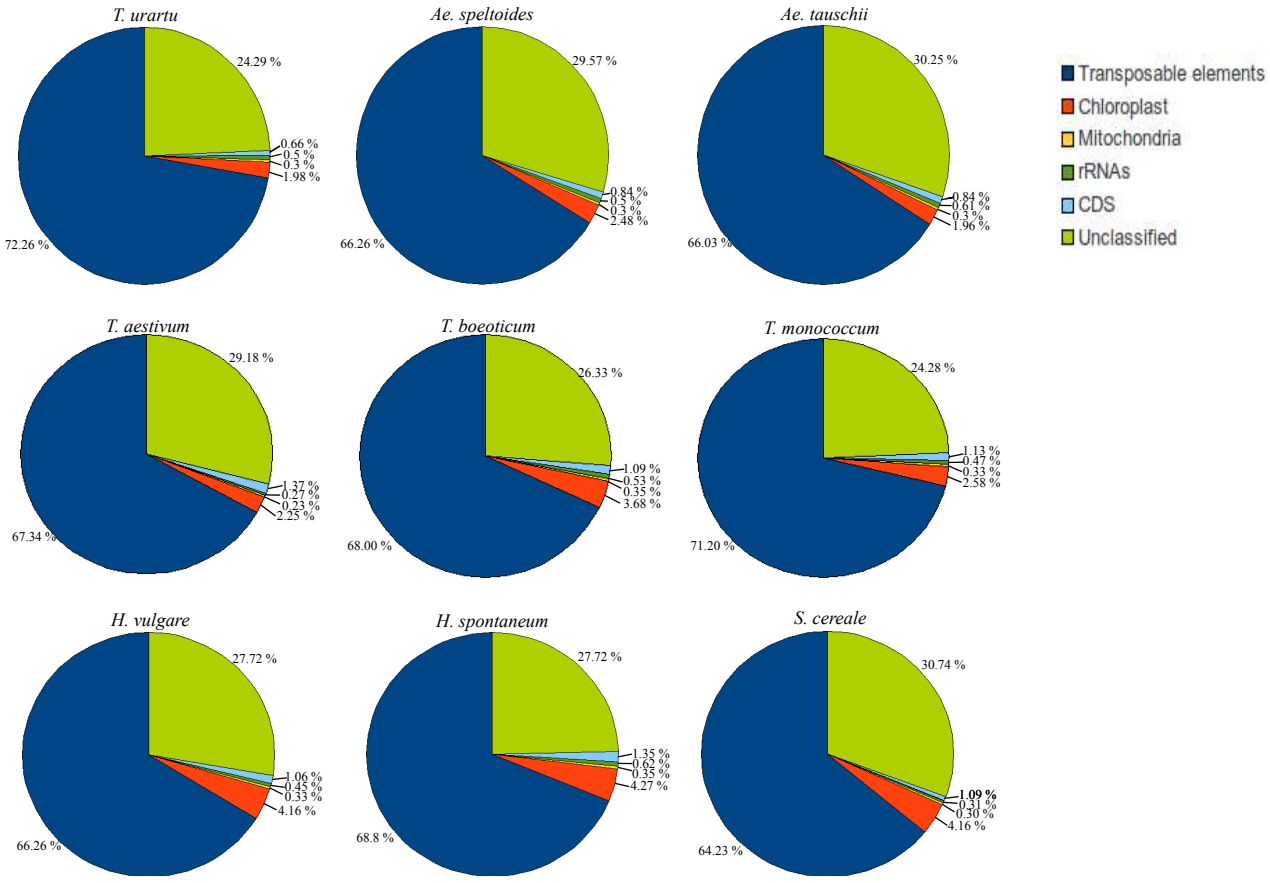


Figure 2. Composition of the 454 reads for each of the individual taxon snapshots. Known TE families make up the largest proportion of the reads. Other sequences identified include chloroplast, mitochondria, rDNAs (ribosomal DNAs); CDS (coding sequences); were also included in the analysis.

Table 1. 454 datasets for each of the 9 species

Taxa name	No. of 454 reads	Average size [bp]	Total [Mb]	Coverage [%]
T. aestivum[a]	499999	415	207	1.2
T. urartu	546057	265	145	2.6
T. monococcum	507523	393	199	3.6
T. boeoticum	458875	540	248	4.5
Ae. speltoides	441540	263	116	2.0
Ae. tauschii	640266	267	171	3.1
S. cereale	586127	292	171	2.9
H. vulgare[b]	1325384	296	392	7.1
H. spontaneum FT11[b]	659263	369	261	4.7
H. spontaneum FT462[b]	642312	375	241	4.4

[a] Sequence supplied by Bevan, M
[b] 2 runs of 454 for *H. vulgare* cv. Barke and 1 run of 454 for *H. spontaneum* FT11 and FT462 supplied by Stein N, and Killian, B

all the taxa analysed, the *BARE1*-clade (which includes *Angela* and *Wis*) made up the largest proportion of the identified TE reads, with between 10.37 and 14.18% classified for all taxa (Figure 3). *Hordeum spontaneum* and *H. vulgare* contained the same proportion of *BARE1* reads with a total of 12.96%, this also confirms previous studies in *H. vulgare* (Vicient *et al.*, 1999; Kalendar *et al.*, 2000; Soleimani *et al.*, 2006; Wicker *et al.*, 2009b) which found that BARE1 contributed to more than 12% of the genome. The hexaploid wheat *T. aestivum* contained a lower percentage of *BARE1* elements (10.54%) than the diploid genome donors, but was similar to *S. cereale* with 10.37%.

The abundance of TE families differs strongly between taxa

The *Gypsy* families *Fatima* and *Sumana* are examples of elements that differ considerably between the barley and wheat species (Figure 3). In *H. spontaneum* and *H. vulgare* only 0.06% of reads were classified as *Fatima* elements. *Secale cereale* also contained a low copy number of *Fatima* elements with 0.45% of reads. This is in contrast to the wheat taxa which

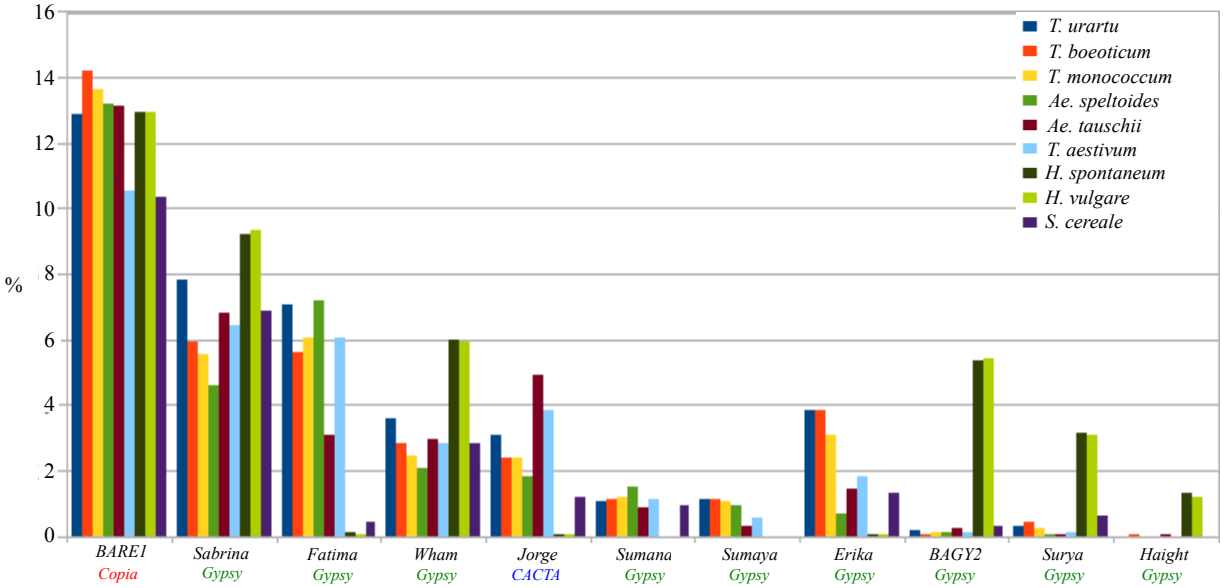


Figure 3. Abundance of TE families and their variation between taxa for 11 transposable element families, displaying the TE families with the largest differences between taxa.

contain between 3.06 and 7.17% *Fatima* elements. The *Sumana* element shows a similar pattern of abundance to *Fatima* between the barley and wheat taxa (Figure 3). The most dramatic differences were observed for *BAGY2*, which was found to make up 5.45% of the overall *H. vulgare* reads and 5.36% of *H. spontaneum* reads. However, *BAGY2* only accounted for between 0.10 and 0.22% in rye and the wheat taxa. The *Gypsy* family *Haight* also occurs in higher abundance in *H. vulgare* and *H. spontaneum* than in the other Triticeae members, with *Haight* accounting for 1.34 and 1.18% in *H. spontaneum* and *H. vulgare*, respectively, but was practically absent from rye and the wheat taxa. No *Sumana* elements were found in *H. vulgare* or *H. spontaneum*, but they were present in increasing abundance of between 0.85 and 1.51% in *Ae. speltoides*, *Ae. tauschii*, the three A genome taxa, *S. cereale* and *T. aestivum*. Similarly, no *Sumaya* elements were identified in either *H. spontaneum*, *H. vulgare* or *S. cereale*, while their contribution to the genomes of *Ae. speltoides*, *T. aestivum*, *Ae. tauschii* and the three A genomes is between 0.27 and 1.14%. *Aegilops tauschii* showed a slightly lower content of both *Sumaya* and *Fatima* elements than *Ae. speltoides*, the A genome taxa and *T. aestivum* (Figure 3).

The *Gypsy* element *Erika*, also displayed a difference in abundance between taxa. The highest numbers of *Erika* were found in the three A genome taxa, where their genomes were made up of between 3.09 and 3.86% *Erika* elements. Whereas the genomes of *T. aestivum*, *Ae. tauschii* and *S. cereale* were made up of approximately half the number of *Erika* elements that were found in the A genome taxa (1.5%). However, in *Ae. speltoides* and the barley taxa low numbers of *Erika* elements were identified, between 0.03 and 1.3%.

Differences in TE classification were not only restricted to retroelements. The *CACTA* element *Jorge* showed very strong variation between taxa, with only 0.03% being attributed to that TE family in *H. vulgare* and *H. spontaneum*. In contrast the abundance of *Jorge* elements increased in *S. cereale* to 1.17%, with a further increase to 1.84% in *Ae. speltoides*. Similar amounts of *Jorge* elements were found in the A genome taxa and *T. aestivum*, which contained between 2.4 and 3.83% respectively. The highest abundance of *Jorge* elements was observed in *Ae. tauschii* which contained 4.93% (Figure 3).

Nucleotide diversity of BARE1 differs between species

As *BARE1* makes up roughly 10–14% of the genomes in all taxa studied, it was possible to assess the nucleotide diversity of *BARE1* elements within each taxon. Nucleotide diversity describes the degree of polymorphism within a population (Nei *et al.*, 1979). In our case, we used it to assess the level of polymorphism of the *BARE1* family within the genome. The *BARE1* clade contains *BARE1* from barley and *Angela* from wheat. Although, 70–80% identical at the DNA level, *BARE1* and *Angela* can be distinguished by some highly diagnostic characteristics (e.g. the *BARE1* LTR starts with TGTT, while *Angela* begins with TGAA). *Angelas* were found to be completely absent from the barley accessions. Rye contains both *Angela* and *BARE1* elements, while wheat contains only minuscule amounts *BARE1*. It was possible to draw phylogenetic trees based on a consensus sequence of the first 300 bp of the LTR of both *BARE1* and *Angela* in all the taxa (Figure 4(a)). The tree clearly shows the close phylogenetic relationship between *BARE1* and *Angela*, with *Angela* arising as a subfamily of *BARE1*.

Nucleotide diversity was tested using the first 300 bp and a region between 600 and 900 bp of the LTR of *BARE1* as queries in BLASTN searches against the 454 datasets. These sequences were chosen as the first 300 bp of the *Angela* and BARE1 LTR evolves rapidly and enables *BARE1* and *Angela* to be distinguished from each other. The second region between 600 and 900 bp was chosen for a second comparison. Regions of 300 bp were selected as the size coincides with the average 454 read length generated for each dataset, so a complete sequence read would match the length of the LTR region selected. For all taxa, we isolated approximately 100 sequences that covered the entire query sequences. The sequences were then aligned using CLUSTALW and nucleotide diversity was calculated from this alignment (Figure 4).

Hordeum vulgare, *H. spontaneum accession* FT11, *H. spontaneum* accession FT462 and *S. cereale* were compared directly with *BARE1*. The nucleotide diversity for the three *Hordeum* accessions is relatively low, with *H. vulgare* scoring less than 0.05 for both LTR regions used for the analysis (Figure 4(b)). Nucleotide diversity for the two *H. spontaneum* genotypes was approximately half that of domesticated *H. vulgare* cv. Barke. In contrast, the nucleotide diversity of *BARE1* elements in *S. cereale* is approximately twice that of domesticated *H. vulgare*, with a diversity of approximately 0.11. (Figure 4(b)).

For the taxa *T. urartu*, *Ae. speltoides*, *Ae. tauschii*, *T. boeoticum*, *T. monococcum* and *T. aestivum* we measured nucleotide diversity of the *Angela* element which is the wheat homologue of *BARE1* (see above). *Secale cereale* was also included in the analysis of *Angela* as it contains both *Angela* and *BARE1* elements. The same LTR regions of the *Angela* element as for *BARE1* were used for the purposes of the analysis. Generally, the *Angelas* in the wheat taxa have higher nucleotide diversities than the *BARE1* element in the three barley taxa, with *Ae. tauschii* having the lowest (0.053) and *Ae. speltoides* the highest (0.076) (Figure 4(c)). However, a lower nucleotide diversity of the *Angela* element was found in *T. boeoticum* and *T. monococcum*, with values of 0.037 and 0.012, respectively (Figure 4(c)). These lower values in *T. boeoticum* and *T. monococcum* are similar to the results obtained for *BARE1* in the three barley accessions. DnaSP ((Librado *et al.*, 2009) was used to assess the statistical validity of

the nucleotide variation within the taxon using Tajima's test (Tajima, 1989) and it was found that all the nucleotide diversity tests were above the 95% confidence level.

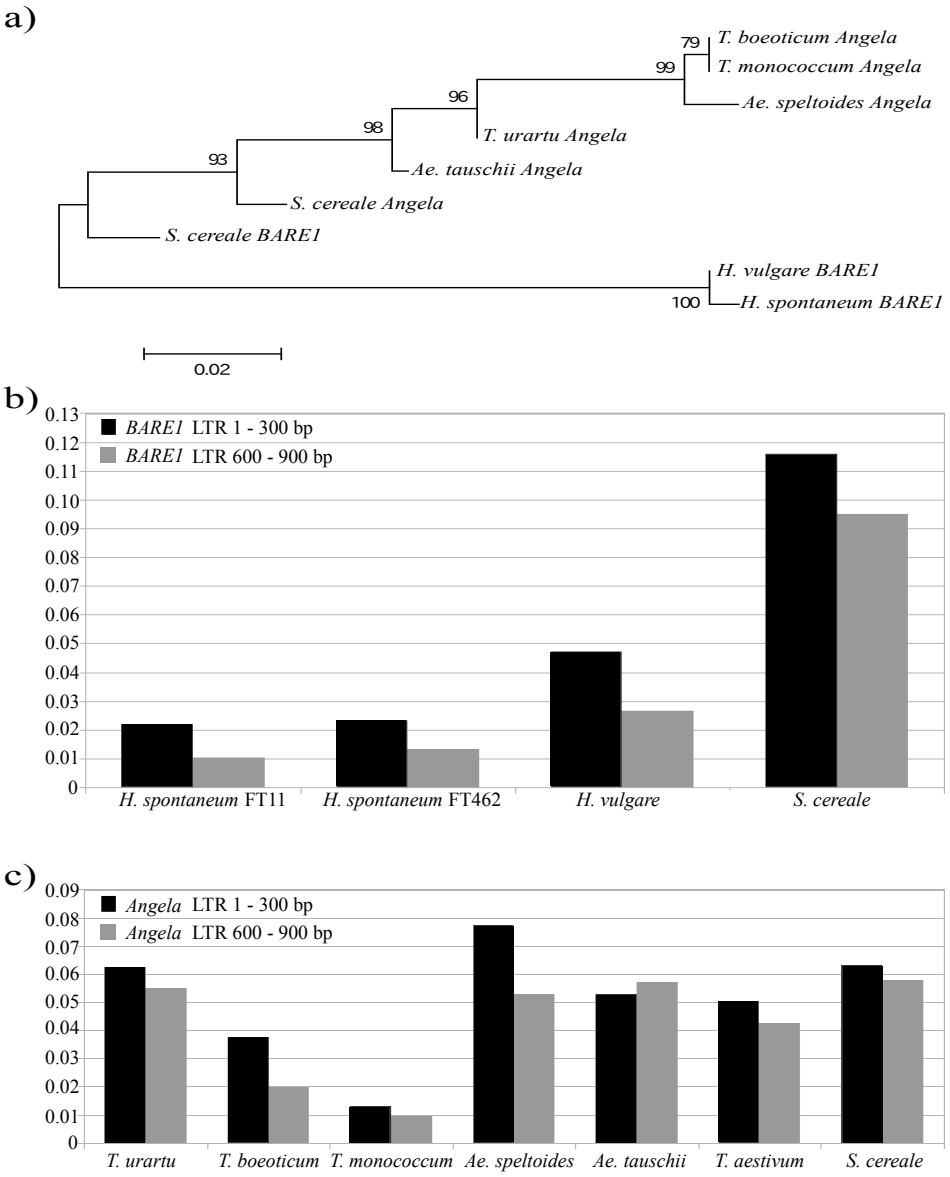


Figure 4. Phylogenetic tree of the consensus sequence of the first 300 bp of the LTR of *BARE1* and *Angela* elements and the nucleotide diversity of two regions of a retrotransposon LTR, using multiple sequence alignment. a) phylogenetic tree showing that *Angela* arose as a subfamily of *BARE1* in the wheat/rye lineage, the tree was produced using MEGA 5.0 with 1000 bootstrap replicates. b) Nucleotide diversity of the barley *BARE1* element with comparison being made between the domesticated *H. vulgare* cv. Barke and the two wild genotypes of *H. spontaneum*. c) Differences in nucleotide diversity between the Triticeae species, showing a difference in the nucleotide diversity between two regions of the *Angela* element LTR.

3.4 Discussion

The objectives of this study were to analyse the genome composition of several Triticeae species, with respect to TE families. 454 titanium sequencing was used to generate between 2 and 5% coverage of the respective genomes. This allowed quantitative comparisons of overall genome composition, providing insight into how these genomes have formed and evolved rapidly in a relatively short evolutionary time of between 10 and 15 million years since they diverged from a common ancestor (Huang *et al.*, 2002a; Akhunov *et al.*, 2003; Chalupska *et al.*, 2008). Approximately 70% of all 454 reads could be classified as known TE families. These results closely mirror previous work by Wicker *et al*. (2009) in which 454 reads from *H. vulgare* were characterised and 69.14% of the reads were found to be TE related. Although coding sequences of genes made up only approximately 1% of the genomic samples, we still sampled approximately 5000 genes per taxon due to the very large number of reads yielded by the Roche/454 technology. Since exploration of gene space was not the focus of this study, a detailed analysis of genic sequences may be presented elsewhere.

The number of reads that remained unclassified ranged from 24.29 to 30.25%; these are more than likely due to unclassified TEs as the predicted levels of TEs in Triticeae are at least 80% (Charles *et al.*, 2008; Choulet *et al.*, 2010). This idea is further reinforced when particular datasets are examined further, for example for *T. urartu* and *T. monococcum* in which the greatest number, 72.26% and 71.20% respectively, of TE families were identified. One possible explanation of this is that the TREP database was originally built with sequences from *T. monococcum*. Thus the TE variety of the A genome is particularly well covered in this database.

The TE composition differs strongly between Triticeae species

Only a few studies have been conducted on TE abundance in Triticeae. Charles *et al*. (2008) and Wicker *et al*. (2009) reported differential amplification of TE families in the A and B genomes, as well as between barley and wheat and our broad dataset shows this to be a general pattern across Triticeae species. We found that some TE families have strongly

varying levels of proliferation in the different taxa. Although, they are found in all Triticeae species analysed, several TE families have undergone either proliferation or a reduction in abundance during species diversification. In some cases, the abundance of TE families reflects phylogeny. For example in barley both *Jorge* and *Fatima* are virtually absent. In rye, they contribute 0.5–1% to the genome while in wheat and its close relatives they contribute a considerable portion of the genome (2–7%). This indicates that these two TE families started to proliferate in the period 2–6 Ma (Chalupska *et al*., 2008) after the common ancestor of rye and wheat diverged from the ancestor of barley. Eventually, *Jorge* and *Fatima* became very successful genome colonisers in the lineage leading to wheat. The *Gypsy* family *Sumaya* shows even more recent pattern of proliferation, as it is present only in wheat and its close relatives but absent from rye and barley.

Another *Gypsy* element, *Erika*, was identified as being present in higher abundance in the wheat taxa containing the A genome, (*T. urartu*, *T. boeoticum* and *T. monococcum*), with it accounting for more than 3% of the overall genome composition. Approximately half of this amount of the *Erika* element was classified in *T. aestivum*, *Ae. tauschii* and *S. cereale*, and *Erika* was found to be virtually absent in *Ae. speltoides*, *H. spontaneum* and *H. vulgare*. This indicates that the *Erika* element proliferated specifically in the A genome lineage. There are also examples of where TE families have proliferated in the lineage leading to barley, while they became virtually extinct in the wheat/rye lineage: The *Gypsy* families *BAGY2*, *Haight* and *Surya* contribute approximately 5% to the genomes of *H. spontaneum* and *H. vulgare*. This is in strong contrast to the rye and wheat species where the three contribute less than 0.5% to the genome.

The question of why some TE families expand their numbers whereas some reduce cannot be answered conclusively with the available data. A possible explanation is that stochastic processes determine the abundance of individual TE families in genomes. Previous studies showed that TEs are active in waves and that the host genome needs time to adapt to newly active TEs by establishing means to silence them (Wicker *et al.*, 2007; Choulet *et al.*, 2010; Slotkin, 2010). We propose that the size of a TE family depends on the level of activity of the

TE and the time needed by the host to establish silencing. In addition, repetitive sequences are rapidly (in evolutionary terms) deleted from the genome through processes such as illegitimate recombination, leading to a constant turnover of intergenic sequences (Devos *et al*., 2002; Wicker *et al*.,2007). Thus only minor variations in these factors could lead to very different TE family sizes between species.

Another question is where do novel TE families come from? We do not believe that any of the TE families studied arose completely de novo in a particular evolutionary lineage. The different TE families differ strongly from each other at the DNA level, indicating that they diverged long before the divergence of the Triticeae species. Thus, the time needed for a new family to emerge by far exceeds the evolutionary time-scale that is studied here. We can also exclude horizontal transfer of TE families from taxa outside the Triticeae because almost all TE families are found at at least a very low abundance in all species examined. We conclude that the vast majority of TE families are actually present in all Triticeae species, but most are present at such low copy numbers that they are hardly detectable in our samples of limited sample size (2–5% of a genome equivalent).

Sequence diversity of BARE1 populations

The large whole-genome samples also allowed an assessment of nucleotide diversity of *BARE1/Angela* elements, the most abundant TE family within Triticeae genomes. Previous studies showed a close phylogenetic relationship of *BARE1* and *Angela* (Wicker *et al.*, 2007; Choulet *et al.*, 2010; Slotkin, 2010). Our data now clearly show that *Angela* arose as a subfamily of *BARE1* in the wheat/rye lineage (Figure 4(a)). This can be seen in *S. cereale*, as rye contains both the *BARE1* and *Angela* elements whereas *Hordeum* only contains *BARE1* and wheat only *Angela* elements. One can speculate that *Angela* out-competed *BARE1* in the wheat lineage, and this could explain the virtual absence of *BARE1* in the wheats. Interestingly, *Angela* diversity is very similar in the rye and wheat taxa, with the exception of *T. boeoticum* and *T. monococcum* (see below). This indicates that the full diversity of *Angelas* had evolved earlier, possibly in the common ancestor of rye and wheat. In addition *T. aestivum*

has the same *Angela* diversity as the subgenome donors. This indicates that *Angelas* have not diverged into many new subfamilies in the genome donors, otherwise *T. aestivum* would display a greater nucleotide diversity of *Angela* elements.

Do mating systems, breeding and geographic isolation influence TE diversity? Interestingly, sequence diversity of *BARE1/Angela* elements varies strongly between different taxa. For example, *BARE1* diversity in *S. cereale* is more than five times higher than in wild barley *H. spontaneum*. In fact, our data suggest that inbreeding taxa (e.g. barley or the A-genome species) tend to have a lower *BARE1/Angela* sequence diversity. Previous studies showed that probably only very few TE copies are active, leading to selective amplification of specific subfamilies within a given TE family (Slotkin and Martienssen, 2007; Slotkin, 2010). We therefore propose that inbreeding tends to keep TE diversity within a genome low. In contrast, outbreeding can lead to the recombination of genomes in which different subfamilies of TEs are active, thereby increasing intragenomic TE diversity.

3.5 Experimental procedures

Plant materials and DNA extraction

The following Triticeae accessions were used: *Ae. tauschii* Coss. subsp. strangulata accession AE429 (from Iran), *Ae. speltoides* var. ligustica accession SPE0061 (from Turkey; single seed descended from AE 346-5-1), *T. urartu* accession EP0471 (from Lebanon; single seed descended from ID 388 studied in Kilian *et al*.,2007a), *T. boeoticum* accession 1628 (single seed descended from ID716 studied in Kilian *et al*., 2007) from Turkey and *T. monococcum* accession 2240 (single seed descended from ID 492 in Kilian *et al*., 2007a) from Turkey. 454 sequences for two *H. spontaneum* genotypes (FT11 from Israel and FT462 from Turkey, both single seed descended) and two runs of 454 sequences of *H. vulgare* cv. Barke were included in the analysis. These were all single seed descended through three generations prior to DNA extraction. *Secale cereale* cv. Imperial seeds were also grown for the purpose of the analysis. Three to five grams of leaf material was harvested from each species for DNA extraction. The DNA extraction was conducted using a 1.3 9 cetyl trimethylammonium bromide (CTAB) and dichloromethane: isoamylalcohol (24:1) method (http://www.protocol-online.org/cgi-bin/prot/). The DNA was further purified with a Qiagen DNeasy Kit, starting at step 13 and following the manufacturer's guidelines. Samples were then sent for 454 titanium (454 Life Sciences, http://www.454.com/) 7 kb paired end sequencing at the Functional Genomics Center Zurich and the IPK in Gatersleben.

Analysis of the 454 reads

LINUX systems (open source operating system) were utilised for the analysis of the datasets. The 454 reads were classified using the BLAST program (http://www.ncbi.nim.nih.gov/). For the identification of TEs we used the databases totalTREP10 (http://wheat.pw.usda.gov/ITMI/Repeats/) and PTREP11. Databases were created locally for the chloroplast, mitochondria, rDNAs, tRNAs and *Brachypodium distachyon* coding sequences. Custom Perl scripts were used to analyse each of the read sets, this created two files, one containing BLAST hits (defined as BLAST hits with E-values <10–6). The second file contained the 'no hits' reads, and this file

was used for subsequent BLAST searches against the other databases.

Assessing nucleotide diversity in BARE1 elements

Consensus sequences of the first 300 bp of the LTR sequence taken for all taxa for each of the *BARE1* element (*H. spontaneum*, *H. vulgare* and *S. cereale*) and the *Angela* element (*S. cereale*, *T. urartu*, *Ae. speltoides*, *Ae. tauschii*, *T. boeoticum* and *T. monococcum*). The program MEGA 5.0 was used with 1000 bootstrap replicates to draw a maximum-likelihood tree with the general time-reversible model of the DNA substitution rate (Tamura *et al*., 2011). The first 300 bp and a region between 600 and 900 bp of the *BARE1* LTR consensus sequence were used as queries in BLASTN searches against the 454 reads. To avoid bias due to the different sample sizes, we used exactly 400,000 reads from each of the 454 datasets for the BLAST searches. For *H. vulgare*, and *H. spontaneum* (FT11 and FT462), all matches of 300 bp in size were used for CLUSTALW alignments. Nucleotide diversity was calculated using an original Perl script. DnasP (Librado and Rozas, 2009) was used to carry out Tajima's test for the statistical analysis to validate the results. For *T. aestivum*, *T. urartu*, *Ae. speltoides*, *Ae. tauschii*, *T. boeoticum*, *T. monococcum* and *S. cereale* we used a consensus sequence of *Angela* (the *BARE1* homologue in wheat). Since *Angela* sequences are more diverse in wheat, we extracted all matches longer than 200 bp from the datasets. All of the 454 datasets can be obtained from the authors by request.

Acknowledgements

This research was supported by COST action FA0604 and the Swiss office for Education and research (SBF) grant number 37150503. We would like to thank Frank Blattner, Fedor A. Konovalov and Andreas Graner for their valuable comments and suggestions on the manuscript, Susanne Konig for excellent technical support and 454 sequencing and the members at the functional genomics centre Zurich for 454 sequencing and the University of Liverpool Genome Centre and the University of Bristol for providing the T. aestivum 454 sequences.

4 Dissecting the Triticeae tribe: Analysis of wheat, barley, rye and their relatives

Christopher P Middleton, Natacha Senerchia, Nils Stein, Eduard D Akhunov, Beat Keller, Thomas Wicker[1] and Benjamin Kilian

[1]Corresponding author

Submitted to *PLOS ONE*

4.1 Summary

Using 454 sequencing, we sequenced the chloroplast genomes of 12 Triticeae species, including bread wheat, barley and rye, as well as the diploid progenitors and relatives of bread wheat *Triticum urartu*, *Aegilops speltoides* and *Ae. tauschii*. Two wild tetraploid taxa, *Ae. cylindrica* and *Ae. genticulata*, were also included. Additionally, we incorporated wild einkorn wheat *Triticum boeoticum* and its domesticated form *T. monococcum* and two *Hordeum spontaneum* (wild barley) genotypes. Chloroplast genomes were used for overall sequence comparison, phylogenetic analysis and dating of divergence times. We estimate that barley diverged from rye and wheat approximately 8-9 million years ago (MYA). The genome donors of hexaploid wheat diverged between 2.1 - 2.9 MYA, while rye diverged from *Triticum aestivum* approximately 3-4 MYA, more recently than previously estimated. Interestingly, the A genome taxa *T. boeoticum* and *T. urartu* were estimated to have diverged approximately 570,000 years ago. As these two cannot be crossed, the divergence time estimate also provides an upper limit for the time required for the formation of a species boundary between the two. Furthermore, we conclusively show that the chloroplast genome of hexaploid wheat was contributed by the B genome donor and that this unknown species diverged from *Ae. speltoides* about 980,000 years ago. Additionally, sequence alignments identified a translocation of a chloroplast segment to the nuclear genome which is specific to the rye/wheat lineage. We propose the presented phylogeny and divergence time estimates as a reference framework for future studies on Triticeae.

4.2 Introduction

The tribe of Triticeae is within the subfamily of the Pooideae and comprises between 400-500 species including diploids and polyploids. It includes several major crop species such as, *Hordeum vulgare* (barley), *Secale cereale* (rye) and *Triticum aestivum* (wheat). These species have undergone many changes during the domestication process, with the domesticated taxa being distinct from their wild ancestors (Doebley *et al.*, 2006; Kilian *et al.*, 2009). However, it was through domestication that these taxa became important for agriculture, with wheat becoming one of the most important crop species.

Triticeae include many polyploid species. The most important is bread wheat (*Triticum aestivum*), an allohexaploid, which has three genomes (A, B, and D) and an approximate genome size of 16-17 Gb (Rees *et al.*, 1965; Bennett *et al.*, 1976). The haploid genome sizes of the progenitors are similar to the haploid genome sizes of other Triticeae and are generally between 3,5 - 8,5 Gb (Eilam *et al.*, 2007; Özkan *et al.*, 2010; Bennett *et al.*, 2011). The complete genome complement of *T. aestivum* was formed from the hybridisation of three diploid ancestors. The first hybridisation event was estimated to have occurred 0.20 to 1.3 million years ago, between *T. urartu* (AA) and a yet unidentified B genome species to form the tetraploid *T. dicoccoides* (Huang *et al.*, 2002a; Dvorak *et al.*, 2005). The exact origin of the B genome is still unclear, but a closely related species or an ancestral relative of *Ae. speltoides* (S genome) has been suggested to be the likely donor (Huang *et al.*, 2002a; Dvorak *et al.*, 2005). The D genome was added to the domesticated tetraploid *T. dicoccon* from *Ae. tauschii* approximately 8,000 - 10,000 years ago to form the complete hexaploid genome complement of *T. aestivum* (Kihara, 1944; Feldman *et al.*, 1995; Kilian *et al.*, 2007b; Bordbar *et al.*, 2011).

Other polyploids within the Triticeae include *Aegilops cylindrica* (jointed goatgrass), a tetraploid containing the C and D genomes and *Aegilops geniculata* (ovate goatgrass) which comprises the M and U genomes (Senerchia *et al.*, 2013). The exact phylogenetic relationships of the C, M and U genomes with others (e.g. tha A, B and D genomes) are less than clear. PCR fragment polymorphism analyses have generally placed the U and M genome closer to D than to the A and B genomes (Tsunewaki, 1996). Furthermore, U and M are probably more closely

related to the D genome than C (Tsunewaki, 1996). In the case of *A. cylindrica*, there is evidence for multiple polyploidisation events, with both the C and the D genome donor acting as maternal parent (Caldwell *et al.*, 2004). *A. cylindrica* is of wordwide economic importance as a weed of bread wheat.

Of less agricultural importance is *T. monococcum* (Einkorn wheat), which contains the A genome and was domesticated from its wild progenitor *T. boeoticum*. Both of these taxa are closely related to *T. urartu*, the A genome donor of *T. aestivum* (Johnson *et al.*, 1976; Kilian *et al.*, 2007a). Even though *T. urartu* and *T. boeoticum* are closely related, they cannot be crossed to produce viable offspring, indicating that their phylogenetic distance is large enough to form a species boundary (Johnson *et al.*, 1976). *H. vulgare* is another agriculturally important species and was domesticated from its wild progenitor *H. spontaneum*.

The evolution and the divergence times of several species from Triticeae have been studied: *Sorghum bicolor* (sorghum), diverged approximately 60 million years ago (MYA) (Paterson *et al.*, 2009b,a), *Oryza sativa* (rice) approximately 40-53 MYA, and *Brachypodium distachyon* diverged approximately 32-39 MYA from the Triticeae (Bossolini *et al.*, 2007; Initiative, 2010). Within Triticeae inferred dates of divergence are sometimes vague: *H. vulgare* is estimated to have diverged from rye and wheat 10 -15 MYA (Huang *et al.*, 2002b; Akhunov *et al.*, 2003; Chalupska *et al.*, 2008). *S. cereale* and wheat diverged approximately 5 - 11 MYA, and the ancestral genome donors *T. urartu*, *Ae. speltoides* and *Ae. tauschii* were estimated to have diverged from each other between 2 and 6 MYA. (Huang *et al.*, 2002b; Akhunov *et al.*, 2003; Chalupska *et al.*, 2008). It is important to narrow down these estimates to gain a better understanding of the evolution of the Triticeae. This is particularly the case for the divergence times of *T. urartu*, *Ae. speltoides* and *Ae. tauschii*, which have received little attention, as focus on these species has generally been in their contribution to the *T. aestivum* genome.

The size of the chloroplast genome is usually between 115 and 165 kb (Jansen *et al.*, 2006). The composition of chloroplasts from the Poaceae family is very similar between species and consists of a large single copy region (LSC), which is approximately 80 kb, and a small single copy region (SSC) of approximately 13 kb in length, located between the two inverted repeat

sequences of approximately 20 kb (Ogihara *et al.*, 2002; Saski *et al.*, 2007).

To date the sequences of approximately 230 chloroplast genomes are publically available. Some of these, including *H. vulgare* and *S. bicolor*, have been used in comparative analysis to ascertain phylogenetic relationships between grasses (Saski *et al.*, 2007), while Chaw *et al,.* 2004, used whole chloroplast sequences from twelve taxa to date the divergence time betweeen eudicots and monocots to 140 - 150 MYA. In addition, Nikiforova *et al,.* 2013 used complete chloroplast sequences from 47 apple species, including wild and domesticated species to date the divergence times of the individual species. Advantages of using chloroplasts are that they are non-recombining and haploid and can be treated as a single locus and that they are maternally inherited (Hirosawa *et al.*, 2004; Nock *et al.*, 2010).

The origin of the *T. aestivum* chloroplast genome has been investigated in several studies (Hirosawa *et al.*, 2004; Golovnina *et al.*, 2007). Golovnina *et al,.* 2007 used the chloroplast *matk* gene along with the *trn*L intron sequence from a large number of Triticeae species and found that the *Ae. speltoides* chloroplast genome sequence had the highest similarity to the chloroplast genome sequence of *T. aestivum*. Therefore, they suggested that *Ae. speltoides* was a close relative of the diploid species that donated its chloroplast genome to *T. aestivum*. Previous approaches to establishing phylogenetic relationships have focused on sequencing one or a small sample of genes (Soltis *et al.*, 2004). However, with the rise of high throughput sequencing which provides much larger datasets for each of the species, chloroplast genomes can usually be assembled as a side product of survey sequencing projects (Nock *et al.*, 2010). Here we used 454 sequencing to obtain chloroplast sequences for 12 Triticeae species with a coverage of between approximately 22x and 92x. These included the three A, B and D genome donors of *T. aestivum*. We wanted to address the following questions: (i) What are the divergence times of the species studied? (ii) Which of the sub-genome donors contributed their chloroplasts to the polyploids such as *T. aestivum*, *A. cylindrica* and *A. geniculata*? (iii) How do the chloroplast genomes of individual Triticeae species differ at the DNA level? We estimated the divergence times of of all 12 Triticeae species from each other and found the times to be more recent than previous estimates. Additionally, a close relative of *Ae. spel-*

toides was confirmed as the chloroplast donor of *T. aestivum* and a relative of *Ae. tauschii* was identfied as the possible chloroplast donor to *A. cylindrica*.

4.3 Results

Chloroplast genome assemblies

A single run of 454 titanium 7kb paired end sequencing was conducted on genomic DNA of 11 Triticeae species and subspecies (Table 1) and additional sequences for *T. aestivum* were provided by the University of Bristol. We included two domesticated taxa *H. vulgare* ssp. *vulgare* and *T. monococcum* ssp. *monococcum* and their wild subspecies progenitors *H. vulgare* ssp. *spontaneum* and *T. monococcum* ssp. *boeoticum*. From here on in the text these subspecies will be referred to as *H. spontaneum* and *T. boeoticum* respectively.

As organellar DNA was not excluded in the DNA extraction, between 1.96 % and 4.27 % of the total number of reads for each taxon originated from the chloroplast (Table 2). Chloroplast DNA insertions into the nuclear DNA make up less than 0.01% of genomic DNA (Matsuo *et al.*, 2005; Sheppard *et al.*, 2009; Lloyd *et al.*, 2011) and are therefore not interfering with the chloroplast sequence assemblies. Due to the relatively small size of the chloroplast genome ($\approx$150 kb), the large number of reads gave a high coverage for each of the chloroplast genomes of 20 to 90 fold (Table 1). This allowed for high quality assemblies of all of the chloroplast genomes because at such high sequence coverage, the number of sequencing errors in the final assembly is negligible (Schatz, 2012). However, the assembly was hampered by the inverted repeat sequence (IR), which could not be resolved into two separate copies with the available sequences and would have required specific and time consuming laboratory procedures. Thus only single-copy regions and one unit of the IR are contained in our chloroplast genome assemblies (Figure 1).

Origin of chloroplasts in polyploid Triticeae species

We first wanted to study if it were possible to conclusively establish the origin of the *T. aes-*

Table 2. Chloroplast assembly information for all twelve Triticeae taxa

Name[a]	Total reads	Cp reads[b]	Cp reads[%][c]	Coverage x	Chl Size [Bp][d]
T. aestivum	499,999	7,972	1.59	25	135,509
T. urartu	546,057	10,825	1.98	22	135,945
Ae. speltoides	441,540	10,934	2.48	24	134,865
Ae. tauschii	640,266	12,564	1.96	26	134,268
Ae. cylindrica	667,485	26,672	4.00	35	133,444
Ae. geniculata	646,327	25,842	4.00	32	137,231
T. boeoticum	458,875	16,868	3.68	40	134,928
T. monococcum	507,523	13,093	2.58	37	136,923
S. cereale	586,127	18,783	3.20	43	135,604
H. vulgare cv. Barke	132,5384	20,133	1.52	86	135,802
H. spontaneum (FT11)	659,263	28,145	4.27	87	135,549
H. spontaneum (FT462)	642,312	31,158	4.85	92	135,864

[a] Taxon name
[b] Number of 454 reads mapping on the chloroplast
[c] Proportion of 454 reads mapping on the chloroplast
[d] Chloroplast size: including the second inverted repeat

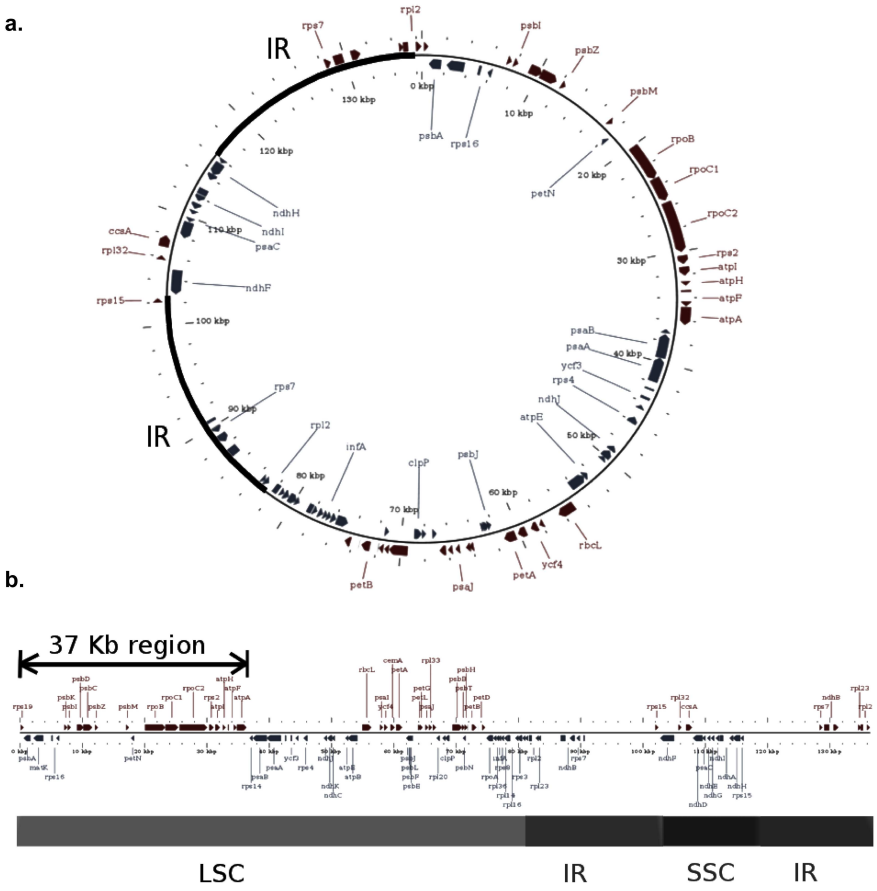


Figure 5. a. Diagram of the layout of the *H. vulgare* chloroplast genome showing the LSC (large single copy) (≈80 kb), the SSC (small single copy) (≈8 kb) and the two inverted repeat (IR) sequences (≈20 kb each). **b.** Shows the sequence assembly, with the arrow representing the 37 kb region chosen for analyses of divergence times.

tivum chloroplast either from *T. urartu*, *Ae. speltoides* or *Ae. tauschii*, the A, B, and D genome donors respectively. A 37 kb sequence from the large single-copy region was chosen for the alignment, because this region is highly conserved between all grass species (allowing reliable sequence alignments) and is not part of the inverted repeat sequence (Figure 5). This region begins in the intergenic sequence 77 bp upstream from the start of the photosystem II protein D1 coding gene (*psbA*) and ends 5 bp before the start of the photosystem I P700 chlorophyll A apoprotein A2 coding gene (*psaB*), (Figure 5).

All 12 taxa were used in multiple sequence alignments of the 37 kb region and 1,000 bootstrap replications were used to draw the phylogenetic tree, using the previously published *B. distachyon* and *O. sativa* chloroplast genome sequences as outgroups (Figure 6). The chloroplasts from the three A genome taxa *T. urartu*, *T. boeoticum* and *T. monococcum* are closely linked in the tree along with the chloroplast from the D genome taxon *Ae. tauschii*. However, it is the chloroplast from *Ae. speltoides* that shows the closest phylogenetic relationship with the chloroplast from *T. aestivum* (Figure 6).

Pairwise analysis was used to determine the sequence similarity between the taxa. The chloroplast sequences of *T. aestivum* and *Ae. speltoides* were found to be 99.87% identical (i.e only 11 polymorphisms in the 37 kb sequence). The other chloroplast genome sequences of the wheat genome donors *T. urartu* and *Ae. tauschii* showed lower sequence identity to *T. aestivum*, ranging from 99.64% - 99.69% and *T. urartu* was found to have a greatest similarity to *Ae. tauschii*, with 99.76%. Additionally, multiple alignments showed many polymorphisms that are only found in the sequences of *T. aestivum* and *Ae .speltoides* that clearly grouped the *T. aestivum* and *Ae. speltoides* chloroplast sequences together (examples are given in figure 6b). Therefore, we conclude that the donor of the *T. aestivum* chloroplast was closely related (but not identical with) today's *Ae. speltoides*.

The chloroplast sequences from the two tetraploid species *Ae. geniculata* and *Ae. cylindrica* were were clustering with chloroplast sequence of *Ae. tauschii* (D genome, Figure 6a). Because the C genome of *Ae. cylindrica* was described to be more distant from D than both U and M (Tsunewaki, 1996; Tsunewaki *et al.*, 2002), our data indicate that the chloroplast of *Ae.*

cylindrica was donated by the D genome parent.

Evidence for the migration of a chloroplast sequence to the nuclear genome

The complete chloroplast genome sequences of *H. vulgare* and *T. aestivum* were directly compared at the sequence level. We identified four deletions and five insertions (InDels) greater than 50 bp in the chloroplast genome sequence of *T. aestivum*, compared to the chloroplast genome sequence of *H. vulgare*.

One sequence of 92 bp is present in the *H. vulgare* chloroplast sequence (position 17,126 - 17,218 bp), but is absent in the same position of the *T. aestivum* chloroplast genome sequence. Interestingly, a homologue of this small region was found on a sequence contig from *T. aestivum* chromosome 3B (accession number FN645450.1) it was found in three locations along the contig in positions 823,371 - 823,445 (74 bp), 839,298 - 839,499 (201 bp) and 924,523 - 924,705 (182 bp), with the sequence from the *H. vulgare* chloroplast being 89% identical to the sequences found on chromosome 3B in *T. aestivum*. Alignments of the three chloroplast sequences revealed that they have been duplicated after insertion into the nuclear genome (Figure 7). We propose that a chloroplast genome segment was copied from the chloroplast genome and inserted into the nuclear genome on chromosome 3B. The segment was then duplicated on chromosome 3B and subsequently and degraded from the chloroplast organellar genome sequence.

All other chloroplast genomes were searched for this deleted sequence, and it was only found in the chloroplast genome sequences of *H. vulgare* and the two wild *H. spontaneum* genotypes. This indicates that this region moved from the chloroplast to the nuclear genome in the lineage leading to rye and wheat after the divergence from barley (Figure 7).

Triticeae divergence time estimates based on chloroplast sequences

Phylogeny of the Triticeae species used in this study was drawn from the chloroplast sequences. Both a maximum likelihood and Bayesian methods were used to obtain divergence time estimates of these species. The maximum likelihood tree was generated using MEGA

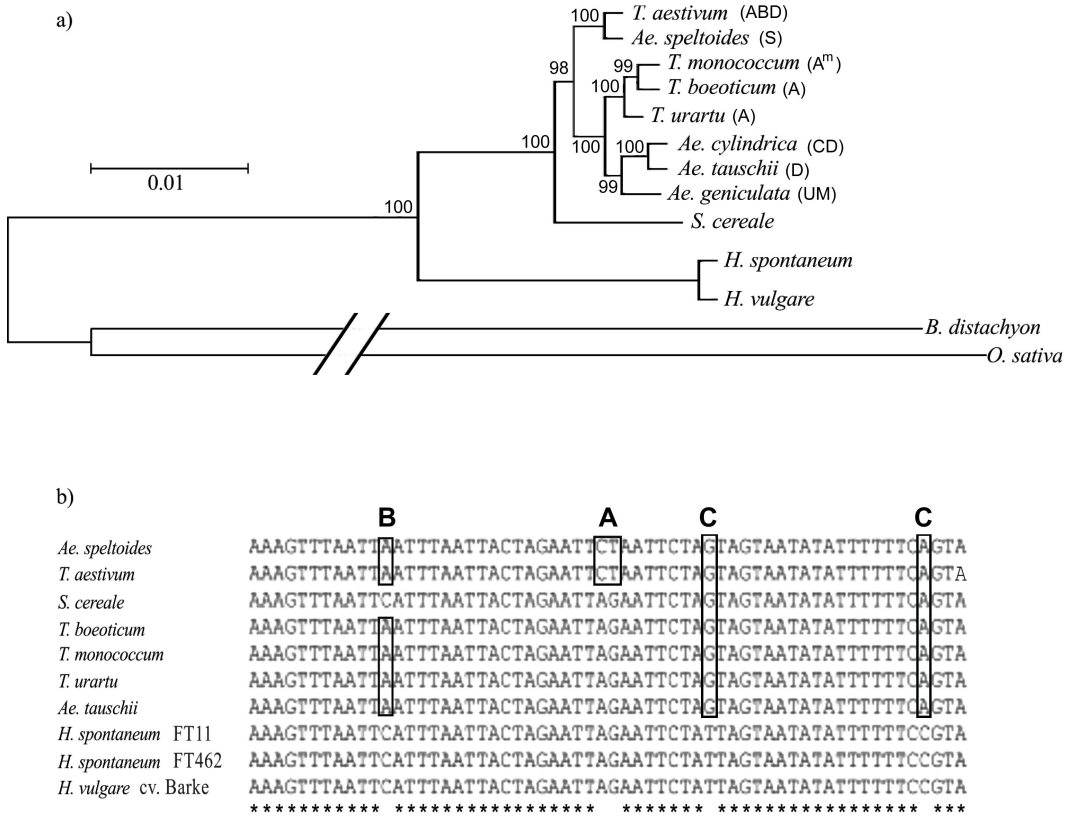
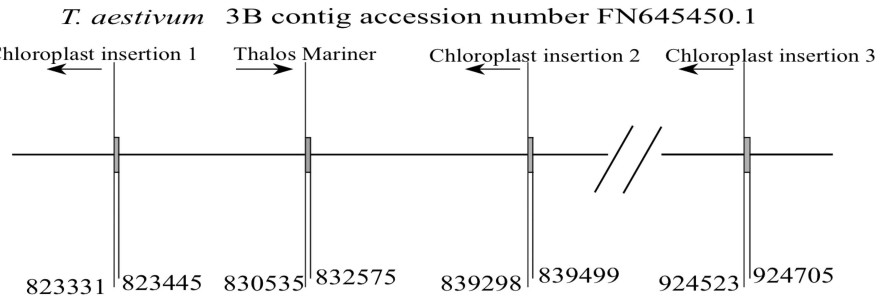


Figure 6. Comparison of Triticeae chloroplast sequences. **a.** Maximum likelihood tree for chloroplast sequences drawn with MEGA 5.0 (Tamura *et al.*, 2011) using 1,000 bootstrap repetitions with *B. distachyon* and *O. sativa* as as outgroups. The genome formula for wheat and its close relatives is indicated in parantheses. It shows that *Ae. speltoides* has the highest sequence similarity to the chloroplast of *T. aestivum*. This indicates that *Ae. speltoides* or a close relative is the chloroplast donor for *T. aestivum*. **b.** Examples of diagnostic polymorphisms. The outlined sequences in (A) are those that are identical between *Ae. speltoides* and *T. aestivum*. (B) indicates nucleotides that are identical amoung the wheats, but different in the other sequences, and (C) indicates positions that distinguish barley from all others.

a.



b.

```
T. aestivum   AAAATAATAGTAAGAATAGACTATAGTAAAAAGTAAAAGGAAA----------------
S. cereale    AAAATAATATTAAGAATAGACTATAATAAAAAGTAAAAGGAAA----------------
H. vulgare    AAAATAATA-TAAGAATAGACTATAGTAAAAAGTAAAAGGAAAAAACCAGCGAATCTTTC
              ********* *************** ******************

T. aestivum   ------------------------------------------------------------
S. cereale    ------------------------------------------------------------
H. vulgare    ATGCTTAAATATCCCGCGAGATGCTTCAAAAGAGAATAAAAATTATTCTAAGAATAAAGA

T. aestivum   ----------------TGGAAAATGTGTGATATGCCGTGAATAGTTCCGTGGAAGAAAGT
S. cereale    ----------------TGGAAAATGTGTGATATGCCGTGAATAGTTCCGTGGAAGAAAGT
H. vulgare    AAGAGTATAAAACAAATGGAAAATGTGTGATATGCCGTGAATAGTTCCGTGGAAGAAAGT
                             ******************************************
```

c.

```
Chloroplast insertion 1 ------------------------------------------------------------
Chloroplast insertion 3 -------------------------ATGGAAGGAAAGAAAAATAATATAAGAATAGACT
Chloroplast insertion 2 GAAGACAATTCTTGACCAAAGAATCAATGGAAGGAAACAAAAATAATATAAGAATAGCCT

Chloroplast insertion 1 ----------------------------------TTTCATGCTTAAATATCCCACGAGGT
Chloroplast insertion 3 ATTGTAAAAGGTAATAGGAAAAAACAAGCAAATATTTCATGCTTAAATATCCCACGAGGT
Chloroplast insertion 2 ATAGTAAAAAGTAATAGAAAAAACCCAGCAAATCTTTCATGCTTAAATATCCCGCGAGGT
                                                  ******************** ******

Chloroplast insertion 1 GCTTCAGAAGAGCATAAAATCCTTCTAAGAATAAGAAAGAAGGTACAAAACAAATGGAAA
Chloroplast insertion 3 GCTTCAGAAGAGCATAAAATCCTTCTAAGAATAAGAAAGAAGGTACAAAACAAATGGAAA
Chloroplast insertion 2 GCTTCAGAAGAGCATAAAATCATTCTAAGAATAAGAAA-AAGGTATAAAACAAATGGAGA
                        ********************** ********** ****** ********** *

Chloroplast insertion 1 TGTGTGATATGCCATGAATAGTTCCGTGG
Chloroplast insertion 3 TGTGTGATATGCCATGAATAGTTCCGTGG
Chloroplast insertion 2 TATGTGATACGTCGTGAATAGTT------
                        * ******* * * *********
```

Figure 7. Putative migration of a chloroplast segment into the nuclear genome in the wheat/rye lineage. **a.** Chromosome 3B section containing a duplicated chloroplast insertion. **b.** A region of the alignment of the chloroplast sequences of *H. vulgare* and *T. aestivum*, showing the removed sequence from the *T. aestivum* chloroplast genome sequence. **c.** Alignment of the three identified chloroplast sequences from chromosome 3B of *T. aestivum* showing that this sequence has undergone a possible duplication event after excision from the chloroplast genome. The highlighted region shows the absent region in the *T. aestivum* chloroplast genome (**b.**).

5.0, with 1000 bootstrap replicates and a GTR+G+I model of substitution. A topologically identical tree was produced using MrBayes under the same substituition model (Figure 8).

To estimate divergence times, we used the 37 kb chloroplast region described above. This region contains 21 genes, 15 tRNAs and seven intergenic sequences greater than 1 kb.

Two distinct methods were also employed to test the validity of the estimates of divergence of the Triticeae used in this project. The first method used a penalised likelihood method assuming a strict molecular clock. Several programs were implemented in order to obtain estimates for the divergence times of the Triticeae species. These included PAUP* (Swafford, 2002) for phylogenetic analysis and tree building, jModeltest (Posada, 2008) to identify the best substitution model of mutation rates as a basis for age estimates and r8s (Sanderson, 2003) to infer divergence times and to test the consistency of the molecular clock along the branches of the phylogenetic tree. We used the previous estimate that *B. distachyon* and *O. sativa* diverged from *T. aestivum* approximately 32-38 MYA and 40-53 MYA respectively (Initiative, 2010), as anchor points for the calculations of the divergence times. Based on the divergence of *O. sativa* and *B. distachyon*, the overall substitution rate for the 37 kb region was calculated to be 1.06E^{-3} per base per million years. From this the divergence time of *H. vulgare* from *T. aestivum* was found to be 8.9 MYA $\pm$ 0.9, (sequence identity of 97.84%). *S. cereale* was found to have diverged from *T. aestivum* 4.0 MYA $\pm$ 0.5 (sequence identity of 99.43%). Both of these estimates of divergence times are more recent and have a smaller deviation than previously published estimates (Huang *et al.*, 2002b; Akhunov *et al.*, 2003; Chalupska *et al.*, 2008). The divergence time of the tetraploid *Ae. geniculata* from D genome species was found to be 1.62 $\pm$ 1 MYA and *Ae. cylindrica* diverged from *Ae. tauschii* approximately 0.18 $\pm$ 0.05 MYA (Figure 8a).

The second method to estimate the divergence times used Bayesian inference as implemented in the software BEAST. The dates of divergence where also based on the calibration points from *O. sativa* and *B. distachyon* described above. The divergence times were calculated twice, using a fixed and an uncorrellated relaxed molecular clock. The results of the relaxed clock analysis are shown in Figure 4b while the comparison of relaxed and strict clock

analyses is shown in Table 2. Because the relaxed clock allows for different subtitution rates in different branches of the tree, the estimates of divergence times have larger confidence intervals than with the strict clock (Table 3). In general, the divergence times derived from the Basyian apporach were very similar to those using a penalised likelihood approach with the Basyian estimates being slightly more recent (Figure 8). The barely/wheat divergence was placed at 8.13 $\pm$2.13 MYA while the divergence of *S. cereale* from wheat was 3.67 $\pm$1.5 MYA. The split in the three clades containing the A, B and D genome donors occurring 2.67 $\pm$1.1 MYA. A further split in the branching resulting in the clades containing the A genome taxa and D genome taxa including *Ae. genticulata* occurring approximately 1.81 $\pm$ 0.8 MYA. The divergence of *Ae. speltoides* from *T. aestivum* was estimated to have occured 0.87 $\pm$ 0.5 MYA (Figure 8b). In addition to using the whole 37 kb interval, we also generated partitions using genes and intergenic sequences separately. The results were virtually identical to those where the sequence was used as a whole.

Table 3. Divergence time estimates used Bayesian inference, applying strict and relaxed molecular clocks

Node[a]	Divergence time[b]	STD[c]	Divergence time[d]	STD[c]
H. vulgare/*H. spontaneum*	0.19	0.17	0.09	0.06
H. vulgare/*T. aestivum*	8.13	2.13	8.19	0.73
S. cereale/*T. aestivum*	3.67	1.46	3.13	0.39
B genome/A+D genomes	2.67	1.10	2.21	0.30
D genome/A genome	1.81	0.79	1.39	0.21
Ae. tauschii/*Ae. geniculata*	1.33	0.62	1.16	0.20
Ae. tauschii/*Ae. cylindrica*	0.34	0.25	0.19	0.09
T. urartu/*T. boeoticum*	0.76	0.45	0.55	0.15
T. boeoticum/*T. monococcum*	0.29	0.22	0.29	0.10
T. aestivum/*Ae. speltoides*	0.87	0.49	0.78	0.20

[a] Node of the phylogenetic tree, see also Figure 8
[b] Divergence time estimate im million years, using a relaxed clock
[c] Standard deviation
[d] Divergence time estimate im million years, using a strict clock

Divergence time calculations of the subspecies

It was also possible to date the divergence of *T. boeoticum* from *T. urartu* to approximately 0.57 $\pm$ 0.14 MYA using the penalised strict clock method and 0.76$\pm$0.45, using the relaxed clock method. The divergence time between *T. boeoticum* and *T. monococcum* could not be calculated using the semi penalised likelihood estimation. However, this was resolved using Basyian inference, with the estimated time of divergence being 0.29 $\pm$ 0.22 MYA (using a relaxed clock) and 0.29 $\pm$ 0.2 MYA (using a strict clock, Table 3).

H. vulgare and *H. spontaneum* are very closely related (sequence identity of 99.98%). The divergence time of the two was calculated to be 80,000 $\pm$ 20,000 years using semi penalised likelihood. Using the Bayesian approach with a strict clock yielded a very similar number of 90,000 $\pm$ 60,000 years and approximately double this with a relaxed clock, 190,000$\pm$ 170,000 MYA.

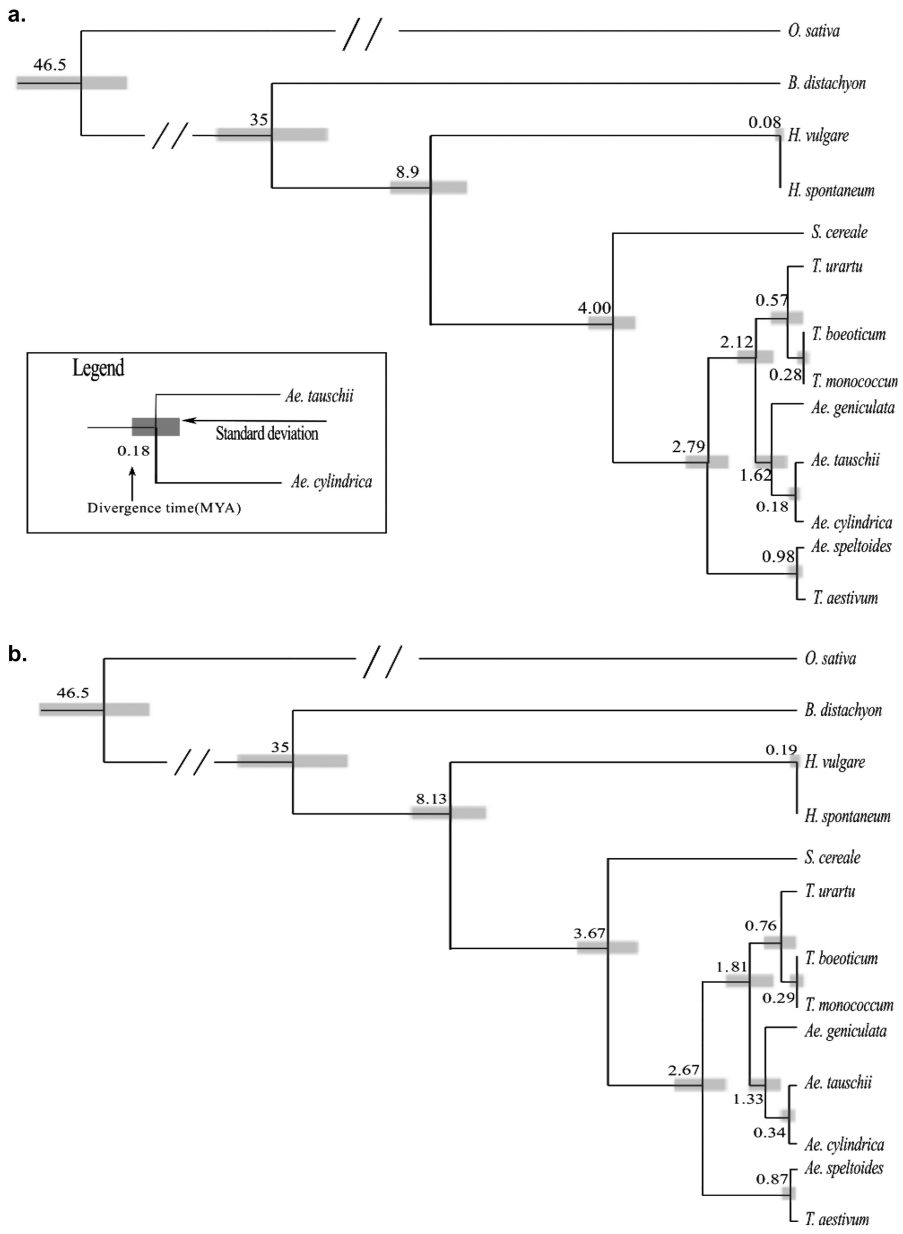


Figure 8. Divergence time estimates of Triticeae species based on chloroplast sequences. **a.** Divergence times of the species based on the strict clock method, using a substitution rate of $1.06e^{-3}$ per base per million years calculated using PAUP and r8s programs over the selected 37 kb region of the chloroplast genome. Divergence times are represented in million years. **b.** Divergence time estimates based on an uncorrelated relaxed clock method. Both trees were drawn using the divergence of *O. sativa* and *B. distachyon* as anchor points. The legend describes the divergence time in million years and the grey boxes represent the standard deviation of the divergence times. The precise numbers for standard deviations are given in Table 2.

4.4 Discussion

The objectives of this study were to examine the relationships of Triticeae species using chloroplast genome sequences. Because the 454 titanium sequencing generated on average 500,000 reads from genomic DNA per taxa, it resulted in approximately 8,000-30,000 reads that were derived from the chloroplast. This led to high quality assemblies of the chloroplast genomes from all taxa, and allowed us to calculate the divergence times of several Triticeae taxa and draw conclusions on the origin of chloroplasts in polyploid species.

Barley, rye and wheat diverged within the past 8-9 million years

The central aim of our study was to obtain more precise estimates for the divergence times of Triticeae species. For these estimates, we used a 37 kb segment of chloroplast sequence which contained both protein-coding and non-coding regions. Most previous divergence time estimates were based on intergenic sequences or synonymous sites of coding regions (Huang *et al.*, 2002b; Dvorak *et al.*, 2005; Chalupska *et al.*, 2008). We argue that it is legitimate to use large segments that contain both intergenic and genic sequences. We partitioned the sequence into coding and non-coding datasets. The results were virtually identical to those obtained by using the 37 sequence as a whole. Furthermore, using strict clock and relaxed clock methods led to almost identical results on the partitioned and non-partitioned datasets. Our divergence time estimates are generally more recent than estimates from previous studies (Huang *et al.*, 2002b; Dvorak *et al.*, 2005; Chalupska *et al.*, 2008). However, due to limited availability of genomic sequences, previous estimates were based on single or very few gene sequences which consequently lead to relatively large standard deviations for the estimates. In particular the divergence of *H. vulgare* from wheat which was estimated previously to have occurred approximately 8-12 MYA (Huang *et al.*, 2002b; Dvorak *et al.*, 2005; Chalupska *et al.*, 2008) is shifted to more recent 8.1 and 8.9 MYA using strinct and relaxed clock methods, respectively. These values are still in the range of estimates reported by Chalupska *et al,.* (Chalupska *et al.*, 2008) (approx. 7-16 MYA) Dvorak *et al,.* (Dvorak *et al.*, 2005) (8.3-11.3 MYA) but clearly more recent than the range of 11.4 $\pm$ 0.6 reported by Huang *et al,.* (Huang

et al., 2002b).

Similarly, our estimate of 3.5-4 MYA for the divergence of rye from wheat and its relatives is more recent than previous ones (Huang *et al.*, 2002b; Chalupska *et al.*, 2008). Again, this value is still in the range of estimates reported by Chalupska *et al,.* 2008 but more recent than the values reported by Huang *et al,.* (Huang *et al.*, 2002b) which were 3-9 MYA and 7.4±0.9 MYA, respectively.

Phylogeny and divergence of wheat and its genome donors

Of particular interest to us was the precise phylogeny and dating of the age of the wheat genome donors. Our analysis indicates that the donors (or their close relatives) of the wheat genomes diverged within the past 3 million years. According to our phylogenetic analysis *Ae. speltoides* branched off first, followed relatively soon by the divergence of *Ae. tauschii* and *T. urartu*. Our estimates are in the range of previous ones (Dvorak *et al.*, 2005; Chalupska *et al.*, 2008) which ranged rather widely from 2-6 MYA. However, we have narrowed this range of divergence to 1.9-2.5 (using a fixed clock) and 1.5 - 3.7 MYA (using a relaxed clock).

The precise phylogenetic relationships between the A, B and D genomes are still a topic of debate. Depending on which and how many gene loci were studied, the A, B or D genome were each once found to be the most divergent of the three (Petersen *et al.*, 2006; Escobar *et al.*, 2011). Reticulate evolution (i.e. hybridisation of closely related species) and incomplete lineage sorting in large populations were proposed as possible explanation for these contradictory results (Escobar *et al.*, 2011). We also consider it possible that, due to the relatively small datasets, ancient paralogs were compared instead of true orthologs. Furthermore, several studies showed that genomic sequences of Triticeae species are composed of haplotype segments that are older than the species that were compared (Isidore *et al.*, 2005; Scherrer *et al.*, 2005; Wicker *et al.*, 2009a). Thus, we must emphasize that our data only provides the phylogenetic relationships of the chloroplast lineages. This is a limitation of our approach, as new chloroplast lineages may have been introgressed independently of the actual species divergence. A general conclusion will probably only be possible once large portions of the A,

B and D genomes are compared.

Formation of a species boundary within the last 550,000-760,000 years in A genome species

Dating of three diploid Triticeae taxa containing the A genome was also conducted, these included *T. urartu*, *T. boeoticum* and *T. monococcum*. The three taxa were all found to have diverged relatively recently, with *T. urartu* diverging from the other two roughly 550,000-760,000 years ago (depending on which estimate is used, see Figure 8 and Table 3). The close relationship between *T. urartu* and *T. boeoticum* is of particular interest because it may give some indication for the time it takes to evolve a species boundary in the Triticeae tribe: these two taxa are not interfertile (Johnson *et al.*, 1976), indicating that a species boundary evolved in less than 550,000-760,000 years. Here, it has to be noted that this estimate refers to the divergence of the chloroplast lineages. Thus, the actual species divergence could be even more recent (discussed below).

In contrast, successful crosses can be made between the very closely related *T. boeoticum* and *T. monococcum* (Kilian *et al.*, 2007a), indicating that both taxa are fully interfertile. Indeed, it was proposed based on archeobotanical findings that *T. monococcum* (the domesticated form of wild einkorn wheat) originated only within the last 12,000 years (Kilian *et al.*, 2007a). However, we have dated the chloroplast divergence between the two to about 280,000-290,000 years. This discrepancy can be explained by the intrinsic characteristics of molecular dating (see dicussion below).

Chloroplasts of cultivated and wild barley are very closely related

Comparison of the *H. spontaneum* (FT11) chloroplast sequence with *H. vulgare* cv. Barke showed that these two sequences are virtually identical with a sequence homology of 99.98%. The polymorphisms were distributed more or less evenly, so we excluded the possibility that one single event could be responsible for the difference (e.g. a micro-rearrangement that affected a dozen or so bp). This high level of sequence similarity translates into a divergence

time of approximately 80,000 ±20,000 years under the strict clock assumption and approximately twice this, 190,000±170,000 years using a relaxed clock approach. This estimate was based on the *H. spontaneum* FT11 genotype from Israel. A second *H. spontaneum* genotype was also included in the analysis (FT462 genotype from Turkey), was found to have a virtually identical sequence to the FT11 genotype (99.98%). From this data we cannot infer the origin of *H. vulgare* cv. Barke from either of the two *H. spontaneum* accessions, as the relationship between these accessions is too close. Nevertheless, we can state that chloroplasts from cultivated and wild barley are clearly more closely related than those of other pairs of wild and domesticated Triticeae subspecies (e.g. *T. boeoticum* and *T. monococcum*, see above). The fact that our estimates for the divergence of wild and cultivated barley predate the beginning of agriculture approximately 10,000 years ago may also be explained by the characteristics of molecular dating (see dicussion below).

The *T. aestivum* chloroplast diverged from that of *Ae. speltoides* less than 1 million years ago

The high quality assemblies allowed us to conclusively determine the origin of the *T. aestivum* chloroplast genome donor. Previous studies by Tsunewaki *et al.* (1983), Provan *et al.* (2004), Kilian *et al.* (2007b) suggested a link between the *Ae. speltoides* chloroplast genome and the chloroplast genome of *T. aestivum*, but these studies were based only on a small region of the chloroplast genome or on RFLP markers. The large 37 kb sequence used in our analysis showed clearly that the *T. aestivum* chloroplast is most closely related to the one of *Ae. speltoides*. In addition to the overall higher level of sequence homology, an abundance of diagnostic nucleotide substitutions demonstrated the close relationship between the *T. aestivum* and *Ae. speltoides* chloroplast sequences.

Additionally, our data allowed us to estimate that the B genome donor diverged from *Ae. speltoides* approximately 780,000-980,000 years ago (again depending oth the estimate used). Because *Ae. speltoides* is a strong outbreeder (Kilian *et al.*, 2007b), a large number of haplotypes may exist and further sampling of *Ae. speltoides* accessions could lead to the discovery

of the same combination of haplotypes that the B genome of *T. aestivum* contains.

Goatgrass chloroplasts are closely related to those of D genome species

The phylogenic tree involving the two tetraploid taxa *Ae. geniculata* and *Ae. cylindrica* showed that their chloroplast genomes are closest to that of the D genome species *Ae. tauschii*. Because both, the U and M genomes were placed very near the D genome in previous studies (Tsunewaki, 1996), we can not draw any conclusions as to which of the two contributed the chloroplast in the polyploidisation event that led to the formation of *Ae. geniculata*. In contrast, some conclusions are possible for *Ae. cylindrica* which contains the C and D genomes: The Triticeae C genome was described to be more distant from D than both U and M (Tsunewaki *et al.*, 2002). Thus, our data indicate that the chloroplast of *Ae. cylindrica* was donated by the D genome parent. Previous studies suggested that multiple independent polyploidisation events have led to the formation of *A. cylindrica* species, with both the C and the D genome donors acting as maternal parents (Caldwell *et al.*, 2004). We therefore conclude that the *A. cylindrica* accession (TA 2204 = AE 719) used in this study represents a lineage where the D genome donor was the maternal parent. Furthermore, we can state that the chloroplasts of the D genome and the two goatgrass species diverged only within the past 1.1 to 1.6 Myr. Precise relationships and origins of the individual chloroplast donors, however, can only be determined once chloroplasts from diploid C, U and M genome species are sequenced.

The advantages and pitfalls of chloroplast dating

Here we used a large part of the chloroplast genome to determine the divergence times of several Triticeae species. Chloroplast sequences have been previously used to determine the divergence time of monocots and dicots (Chaw *et al.*, 2004) as the evolutionary distance between them is relatively large (approximately 130 MYA). There are examples of where whole chloroplast sequences have been used to measure phylogeny and estimate divergence times of closely related species. This was utilised previously (Nikiforova *et al.*, 2013), where chloroplast genomes from 47 apple taxa were used, including both wild and domesticated taxa. By

using the whole chloroplast sequence, a complete phylogeny and estimates of the divergence times of the taxa could be obtained (Nikiforova *et al.*, 2013).

Robust calibration dates are an important consideration when anchoring the tree for the purpose of dating divergence times. Due to the absence of a fossil record in the Triticeae, the calibration dates had to be taken from more distantly related species. We therefore used as calibration dates the divergence of *O. sativa* and *B. distachyon*, which were estimated to have diverged between 40-53 MYA and 32-39 MYA from the Triticeae, respectively. These estimates were based on the comparison between orthologues gene pairs from rice, Brachypodium and Triticeae species (Initiative, 2010).

The use of chloroplast sequences to estimate species divergence is, in principle, not problematic if the species are separated by a large evolutionary distance. However, with shorter evolutionary times, such as those seen in the Triticeae which are typically less than 10 MYA, the accuraccy of estimating the divergence time based on chloroplast genome sequences decreases. The reason is that the possible presence of multiple haplotypes and/or independent chloroplast lineages does not correspond with the actual species divergence. The example in Figure 9 shows that lineages of chloroplast haplotypes which are present in two species today might have diverged substantially earlier than the actual species. Indeed, this is exactly what was found for mitochondrial DNA which was used to estimate the divergence times within a series of bird species: all divergence time estimates were almost double the divergence dates based on the fossil record (Edwards *et al.*, 2000; Arbogast *et al.*, 2002; Steiper *et al.*, 2008). This highlights a principle problem of molecular dating, namely that all divergence times are over-estimated. This is because various ancient haplotype lineages can be present and be recombined within a population or species. Only after species become reproductively isolated, haplotype lineages can no longer mix. Therefore, molecular dating can only estimate the divergence of a particular genetic locus (in our case the chloroplast) but not the divergence of species (Figure 9). Thus, for example the estimated 600,000-700,000 years for the formation of the species boundary between *T. boeoticum* and *T.urartu* has to be seen as the upper limit. This can also explain the discrepancy between our estimated 270,000-280,000 years

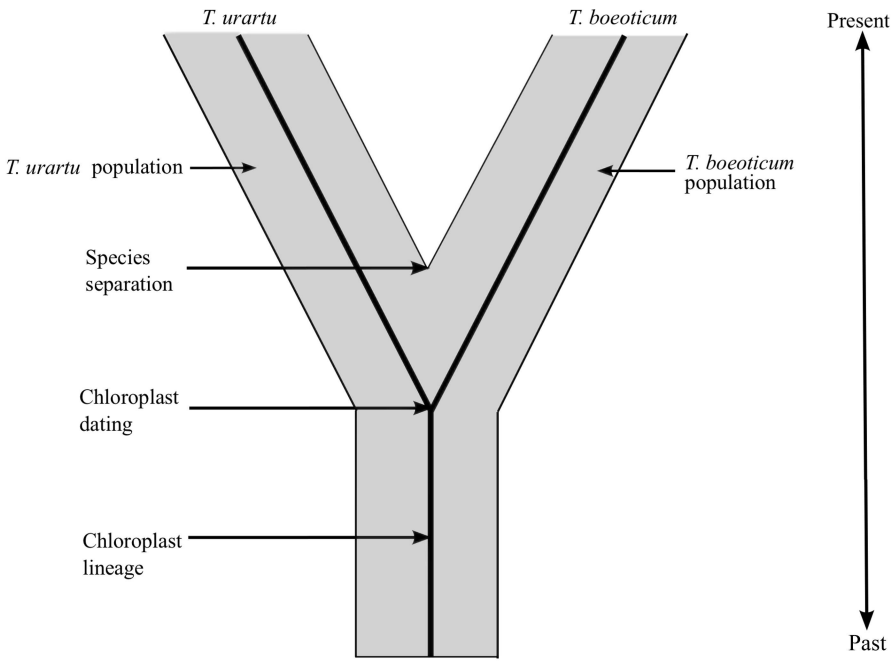


Figure 9. Diagram showing how haplotype divergence and/or incomplete lineage sorting can result in an over-estimation of divergence times, using *T. boeoticum* and *T. urartu* as examples. Haplotypes and/or chloroplast lineages diverge before the formation of the new species (i.e. the complete reproductive isolation of the two populations). Consequently, divergence time estimates derived from sequence data are, in principle, always over-estimates.

for the *T. boeoticum*/*T. monococcum* divergence and the 12,000 year-old archeological evidence: Species have to be morphologically different to be distinguished in the archeological or fossil record. Thus, archeological or fossil evidence must always lead to under-estimated divergence times.

In conclusion and with these limitations in mind, we suggest that the use of a large 37 kb region of the chloroplast genome provided robust and relatively precise divergence time estimates for the main Triticeae species. In particular, we were able to describe relationships and divergence times between wheat and its close relatives in great detail. For future studies, it will be highly interesting to include sequences of other wild tetraploid Triticeae species (e.g. *T. dicoccoides* and *T. araraticum*), to place them on the phylogenetic tree and calculate their divergence times.

4.5 Experimental Procedures

Chloroplast assembly

Chloroplast assembly was conducted using Newbler at default settings (454 Life Sciences, Roche). The complete 454 read dataset for each species was used for the purposes of the assembly. In all cases the largest contigs produced during the assembly process belonged to the chloroplast. The final assembly process involved the use of dot plots against a reference sequence, either *H. vulgare* (accession number NC008590), downloaded from NCBI, for the re-assembly of *H. vulgare* and for the assembly of *H. spontaneum*. The *T. aestivum* (accession number NC002762) chloroplast sequence downloaded from NCBI was used a reference sequence for the assembly of *S. cereale*, *T. urartu*, *Ae. speltoides*, *Ae. tauschii*, *T. boeoticum*, *T. monococcum* and for the re-assembly of *T. aestivum*.

Alignment and phylogenetic analysis

Phylogenetic analysis was carried out using a 37 kb region at the start of the LSC. The previously published chloroplast sequences of *B. distachyon* accession number NC011032 and *O. sativa* accession number NC001320 were used as outgroups to anchor the tree. Alignment of all the 37 Kb sequences was done using SeaView (Gouy *et al.*, 2010). jModeltest (Posada, 2008) was used to obtain the substitution model by implementing the hierarchical likelihood ratio test. According to the Akaike information criterion (AIC) the model best fitting the observed data was GTR+G+I (general time reversible). The maximum likelihood tree was drawn using MEGA 5.0 with 1000 bootstrap replicates.

The model parameters were used in the PAUP* program for the likelihood estimation of the branch lengths of the tree. These branch length estimates of the tree were used to compute the divergence times of all the species from *T. aestivum* using the the semiparametric penalised likelihood method implemented in the r8s program (Sanderson, 2002), the smoothing parameter was also estimated using the method described by (Sanderson, 2002). *O. sativa* was used as the outgroup to root the tree and the outgroup was pruned before the divergence times were estimated. The divergence time calculations were based on the assumption that

B. distachyon and *T. aestivum* diverged between 32 - 39 MYA (Bossolini *et al.*, 2007; Bortiri *et al.*, 2008; Initiative, 2010). 100 Bootstrap replicates were conducted on the dating using fseqboot, which is part of the phylip package, and r8s boot kit to transform the replicates for use in the r8s program.

Calibration and estimating divergence times of Triticeae

Due to the absence of good fossil evidence for the Triticeae, calibration of the nodes in the tree was based upon previous molecular data, using the confidence interavals previously stipulated from these findings. Three nodes were selected for the purposes of calibration and these included the divergence times of *O. sativa* 40 -53 MYA (Initiative, 2010), *B. distachyon* 32 - 39 MYA (Initiative, 2010) and *H. vulgare* 6 - 15 MYA (Chalupska *et al.*, 2008).

Divergence time estimates were estimated using the Baysian method implemented in the BEAST programme (Drummond *et al.*, 2007). This software was used to infer tree topology, branch lengths and nodal ages using Baysian inference and Markov chain Monte Carlo (MCMC) analysis. This was conducted using the whole aligned 37 Kb sequence and a partioned data set containing 12 genes and 2 intergenic sequences. The genes used in the partioned dataset include *atpF*, *atpH*, *atpI*, *matK*, *psbA*, *psbC*, *psbD*, *psbK*, *psbZ*, *rpoB*, *rpoC1*, and *rpoC2* with all of these genes being located within the 37 kb region. The intergenic sequences used in the analysis were between the trnS tRNA and the *psb*D gene, which resulted in a sequence of approximately 1080 Bp and the second intergenic sequence is located between the trnC tRNA and the *rpo*B gene with an approxinmate length of 1140 Bp. The individual gene sequences were aligned and concatanated to produce a total aligned sequence of 19033 Bp. The GTR+G+I substitution model was seleted for the genes *atpF*, *atpH*, *atpI*, *psbC*, *psbZ*, *rpoB*, *rpoC1*, and *rpoC2* in the partition, with four gamma categories, with the HKY+G and four gamma categories substitution model was selected for *matK*, *psbA*, *psbD*, *psbK* and for the two intergenic regions, with an uncorrelated relaxed clock model being used, as this allows for rate variation accross the branches, and a Yule tree prior was used to model speciation. Two independent MCMC runs were performed for 10,000,000 generations and

sampling was conducted every 100th generation. *Brachypodium distachyon* was constrained as the outgroup with a mean of 35 Ma and a standard deviation of 1. Convergence between the runs and the amount of burn in were determined using Tracer 1.5 (Drummond *et al.*, 2007), this was used to assess the effective sample size (ESS) and to check the consistancy of the result. TREEANNOTATOR 1.6.2 (Drummond *et al.*, 2007) was used to calculate a maximum clade probability tree using a posterior probability limit of 0.5, with the final tree being visualised in FIGTREE 1.3.1.

Data deposition

The chloroplast genomes described in this study were deposited at GenBank under the following accession numbers: *Ae. speltoides*: JQ740834, *Ae. tauschii*: JQ754651, *H. vulgare* (cv. Barke): KC912687, *H. spontaneum* (accession FT11): KC912688, *H. spontaneum* (accession FT462): KC912689, *T. monococcum*: KC912690, *S. cereale*: KC912691, *T. boeoticum*: KC912692, *T. urartu*: KC912693, *T. aestivum*: KC912694, *Ae. cylindrica*: KF534489, *Ae. geniculata*: KF534490.

Acknowledgements

This research was supported by COST action FA0604 and the Swiss Office for Education and Research (SBF) grant number 37150503.

5 Transcriptome sequencing of pathogen infected wheat reveals diverse expression patterns of transposable elements

Keywords: transcriptome, transposable elements, powdery midew, septoria

5.1 Summary

Expression patterns were analysed for a number of transposable elements in *T. aestivum* after infection with *Mycosphaerella graminicola* and *Blumeria graminis*. By using RNA sequencing data taken from a number of time points after infection it was possible to study how the infection influences the expression of transposable elements. The most abundant retrotransposons in the wheat genome *Angela*, *Sabrina* and *Fatima* were analysed. Additional, retrotransposons included were *Leolyg* and *Stasy*. Other elements investigated include the MITE *Thalos*, the LINE *Stasy* and the *CACTA* elements *Isaac*, *Caspar* and *Clifford*. Differences in the expression patterns were found, with the *Angela* showing a pattern of expression that would suggest it was active in the genome and the other elements such as *Sabrina* and *Fatima* showing evidence of silencing through RNAi. The varied patterns of expression seen between different transposable elements will require further investigation to fully understand how these elements are induced into transposition or silenced during pathogen infection.

5.2 Introduction

The ascomycete *Mycosphaerella graminicola* causes Septoria tritici blotch and is one of the most important diseases in wheat worldwide. Total losses from Septoria can reach up to 50% in areas where the disease is prevalent (Goodwin, 2007). The commencement of the disease cycle usually starts with wind bourne spores that are found on the stubble remaining from the previous seasons crop, these are then passed onto wheat seedlings, establishing the primary infection (Goodwin, 2007). The life cycle of *M. graminicola* occurs in three distinct stages, with the first being the biotrophic phase that occurs after *M. graminicola* penetrates the plant through the stomata (Cohen *et al.*, 1993). During the first 10 days the pathogen goes through its biotrophic phase, with no visible symptoms being apparent on the leaf tissue. After approximately 10 days the first signs of chloroses starts to develop, indicating a switch to necrotrophic stage of the infection cycle. The necrotrophic phase of infection continues until the infected tissue is dead and the pathogen enters the saprotrophic phase and can survive on the dead plant tissue for several months. The initial symptoms of the disease begin with the formation of small brown lesions of necrotic tissue on the leaf, with these symptoms typically appearing 14-21 days after the initial infection. The exact changes that are involved in both the pathogen and host during infection are not clearly understood. The publication of the complete genome sequence of *M. graminicola* should help to identify the genes and their products that enable the infection cycle (Goodwin *et al.*, 2011). The complete genome sequence of *M. graminicola* reveals that it contains 21 chromosomes, with this including 13 core chromosomes and 8 dispensable chromosomes. The core chromosomes contain approximately twice as many genes per Mb and a much lower amount of repetitive sequences when compared to the disposable chromosomes. The genes on the disposable chromosomes were found to be usually truncated compared to the genes located on the core chromosomes (Goodwin, 2007).

Control of *M. graminicola* disease is usually conducted in two ways, either through the use of fungicides or through the breeding of resistant cultivars. So far 15 genes *stb1-stb15* have been identified and characterised in *T. aestivum* that convey resistance to *M. graminicola* (Ghaffary *et al.*, 2012). However, the changes that occur in the transcriptome of *T. aestivum* under in-

fection of *M. graminicola* have so far not been investigated.

Another pathogen of economic importance is *Blumeria graminis* (powdery mildew) and is an obligate biotroph, with its life cycle of growth and reproduction occurring on living epidermal tissue. Powdery mildew can cause disease on a large number of plants, with these being some of the major crop species such as barley and wheat. *B. graminis* has both a sexual and an asexual life cycle, with the infection of the plant occurring during the asexual phase. Once the fungus has penetrated the plant cell it forms a specialised structure, the haustorium which invaginates the plasma memebrane and enables the fungus to assimilate nutrients from the plant and transfer fungal components into the plant cell. Resistance in wheat is usually through gene for gene interactions between effectors from the fungus and resistance proteins in the plant (Dangl *et al.*, 2001). This has lead to an "arms race" between the fungus and and the plant, with this putting the pathogen under strong diversifying selection to change or discard effectors once they are recognised by the plant. Genome sequences are now available for the both *B. graminis f.sp. hordei* and *B. graminis f.sp. tritici*.

As TEs make up a large proportion of the Triticeae genome, the organism has several mechanisms to deter TE proliferation, such as silencing through siRNAs and methylation (Onodera *et al.*, 2005; Qi *et al.*, 2006). siRNA has previously been described in *Arabidopsis thaliana*, but so far no studies were conducted on the expression of TEs in Triticeae. Furthermore, transposable elements have been known to become active due to abiotic or biotic stress. However, the exact nature of how TEs are expressed under these circumstances has only partially been explained. In this study we will look at expression patterns of TEs when *T. aestivum* is infected with *M. graminicola* and *B. graminis*. We hoped to gain insight into how TEs are expressed during infection and whether we find evidence for siRNA silencing of the elements.

5.3 Results

Transcriptome analysis of wheat infected with *Mycosphaerella*

50 bp Illumina reads for three time points from *T. aestivum* were obtained from leaf tissue infected with *Mycosphaerella graminicola*, which causes Septoria tritici blotch. The tissue was sampled at time intervals of 7, 13 and 56 days post infection. However, time zero was not sampled. The sample times coincide with the infection cycle of *Mycosphaerella graminicola*, as the initial infection begins during the first 12 days of infection and visible symptoms start to develop on susceptible cultivars of *T. aestivum*, between 12 and 15 days post infection. During this period the fungal pathogen switches from its biotrophic phase to a necrotrophic phase of its disease development.

As a completed genome sequence for *T. aestivum* is not available the reads obtained from the RNA sequencing were mapped against the latest release of the TREP database, with the gene mapping being carried out on CLC. RNAseq reads were mapped to approximately 500 individual TEs in the day 7 sample, with this number falling to approximately 450 by day 13 and a further reduction to approximately 100 TEs after 56 days. At 56 days post infection the number of recordable reads that mapped to the TREP database decreased dramatically, when compared to the 13 day post infection sample. The drop in transcripts seen in the 56 day sample, is probably caused by the fungus, having been in the necrotrophic phase for several weeks resulting in the plant tissue dying. However, there are still considerable levels of expression seen.

MITEs show the highest level of expression

In the day 7 and and day 13 samples the element that showed the highest expression is the stowaway miniature inverted repeat transposable element (MITE) *Thalos*, with this element having a RPKM number of 155,859 after 7 days and a RPKM (reads per kilo base per million) value of 153,299.82 after 13 days. Several other elements were highly expressed at both 7 and 13 days these include several other MITES, such as *Icarus* and *Pan*, with a total of nine

MITES being the most active TEs in the top twenty expressed. Long interspersed elements (LINEs) also make up a proportion of the most highly expressed TEs, with the LINE *Stasy*, showing an RPKM value of 17,523.404 at Day 7, a RPKM value of 15,577.555 at day 13, and no expression at day 56. As with all other TEs studied this LINE shows a steady decline in expression when comparing samples taken at Day 7 or Day 13.

Transcripts that relate to two LTR retrotransposons are also present in high numbers. The two LTR retroelements that were studied in detail are the *Copia* elements *Leolyg* and *Susanda*. The LTR element *Leolyg* was found to have 7,706 reads mapping to the 4,598 bp element after 7 days, with the number of reads mapping dropping to 2,706 reads after 13 days and dropping further to just 5 reads that mapped after 56 days. The number of transcripts identified for the *Susanda* element showed a similar pattern of the number of reads that mapped, with 7,706 reads mapped on the 6,433 bp element at Day 7 and this decreasing to 4,115 reads that mapped at day 13, with this being reduced further to just 4 reads that mapped after 56 days.

Three other LTR retrotransposons were also investigated as these were found to be the most abundant elements in the genome of *T. aestivum*. These were the *Copia* element *Angela*, and the *Gypsy* elements *Fatima* and *Sabrina* (Middleton *et al.*, 2013). The *Copia* element *Angela* makes up approximately 12% of the overall genome of *T. aestivum*. However, expression of this element is relatively low, with a RPKM value of 1159.51 at 7 days post infection, with this figure rising slightly to 2161.24 at day 13, while no expression is recorded after 56 days. The *Gypsy* elements *Fatima* and *Sabrina* show a similar pattern of low expression with expression falling after each time point. *Sabrina* showed a drop in expression from 102.27 at day 7 to 74.54 after 13 days and no expression was recorded at day 56. *Fatima* was found to be expressed at all time points with RPKM values of 981.25 at day 7, 705.82 and day 13 and 450.59 at day 56.

Expression patterns along transposable elements

The pattern of where the reads mapped to each element was also examined this was done

by using integrated genome viewer (IGV), this is a software program that allows users to visualise the pattern that reads map either to an individual sequence or a complete genome. By visualising the pattern of that the reads mapped to each transposable element it is possible to gain some understanding into how these elements are either expressed or silenced.

This pattern was looked at for the MITE *Thalos*, the LINE *Stasy* and the LTR retrotransposon *Leolyg*. Very different patterns in the way in which the reads mapped are found in each element. The MITE *Thalos* shows reads that map across the whole element in the sample taken at day 7, this changes when the pattern is viewed after 13 days, with only the first 60% of the element, having reads that map to it. At day 56, although only 12 reads actually map to the element they map to the whole element (Figure 10).

The LINE *Stasy* element showed a similar pattern, between sampling on day 7 and day 13, with the reads mapping the entire length of the element. The higher peaks in the number of reads that mapped is probably due to the reads mapping to the coding regions (Figure 11).

The LTR retrotransposon *Leolyg*, had a strong pattern of reads that mapped to each region, with a large number of reads mapping to the CDS located in the middle of the element at the two time points of day 7 and day 13 (Figure 12). The two LTR sequences, also showed a high number of reads that mapped. However, as these regions are direct repeats it is difficult to pinpoint how many are mapped to each LTR sequence. The LTR sequence is thought to act as a promoter for the expression of LTR retrotransposons, therefore the number of reads could be due to the mechanisms that the cell takes to avoid the proliferation of retrotransposons and could be related to the silencing mechanism.

Transcriptome analysis of wheat infected with powdery mildew

RNA sequencing data were also produced for *T. aestivum* infected with *Blumeria graminis*, samples were taken at five time points after infection, 4 hours. 8 hours, 12 hours, 24 hours and 48 hours. The sequencing resulting from the sampling produced approximately 250 million 50 Bp reads for each of the time points. These reads were then mapped to transposable element database (TREP) and mapped reads onto individual elements were visualised with

IGV. Several elements were analysed, these include the most abundant transposable elements in the wheat genome. The LTR retrotransposon *Angela*, *Sabrina* and *Fatima* as these form approximately 15% of the overall genome (Middleton *et al.*, 2013). Other transposable elements analysed include members of the CACTA family, including *Isaac*, *Clifford* and *Caspar*. As no replicates at each of the time points were taken, no real conclusions about the changes of expression could be drawn, but an insight into the pattern of expression of the elements could be reached.

The *Angela* element shows expression accross the whole whole element, with the expression levels differing in the region of the element. The LTRs and the gag ORF are heavily expressed, with the gag ORF being expressed at least one order of magnitude more than the RT/INT ORF. In contrast to the *Angela* element the *Sabrina* element is only expressed in small regions, with the RT ORF showing virtually no expression, with very short segments being highly expressed, and this could represent small interfering RNAs (siRNAs) that function to silence the element. The *Fatima* element shows another different pattern of expression with the LTRs heavily expressed and the CDS only being expressed in small regions, suggesting the small segments in the CDS represent siRNAs, but the reason for the high LTR expression is unknown (Figure 13).

As in LTR retrotransposons, *CACTA* elements show very diverse expression patterns and also differ in the overall amount of reads that could be mapped to these elements. *Isaac* shows expression in the transposase region as well as downstream of it, suggesting the presence of an additional, yet unidentified ORF. Similarly, *Clifford* shows expression in both ORFs. The strong peak of expression upstream of ORF1 is not understood. In contrast, *Caspar* shows a pattern that is similar to that of *Sabrina* with only very short segments being expressed (Figure 14).

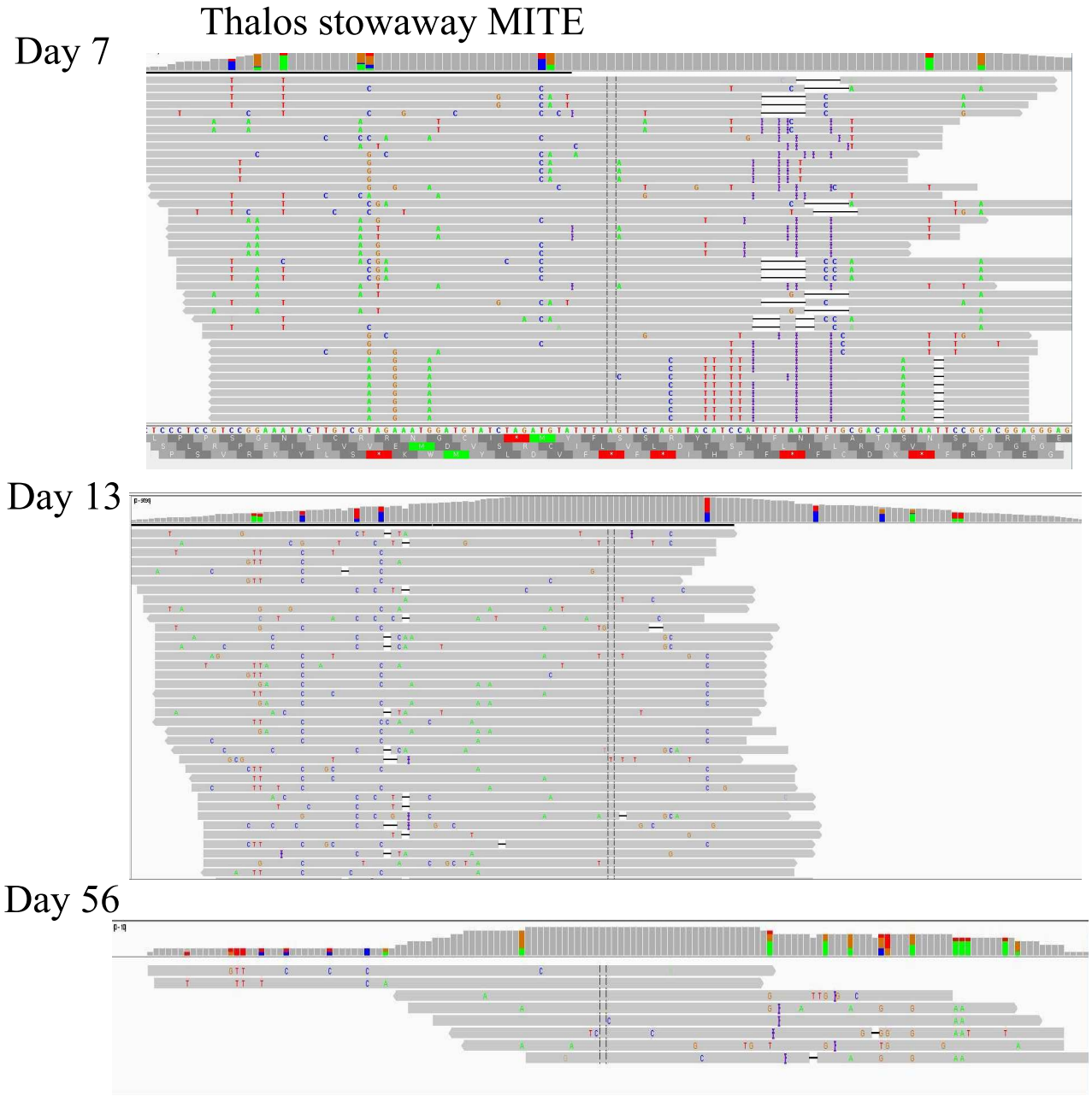


Figure 10. The pattern that the illumina reads map to on the MITE thalos after infection with *Mycosphaerella graminicola*. At day 7 complete coverage is observed, at day 13 only half of the element is expressed and at day 56 the whole element is expressed, but at much lower levels.

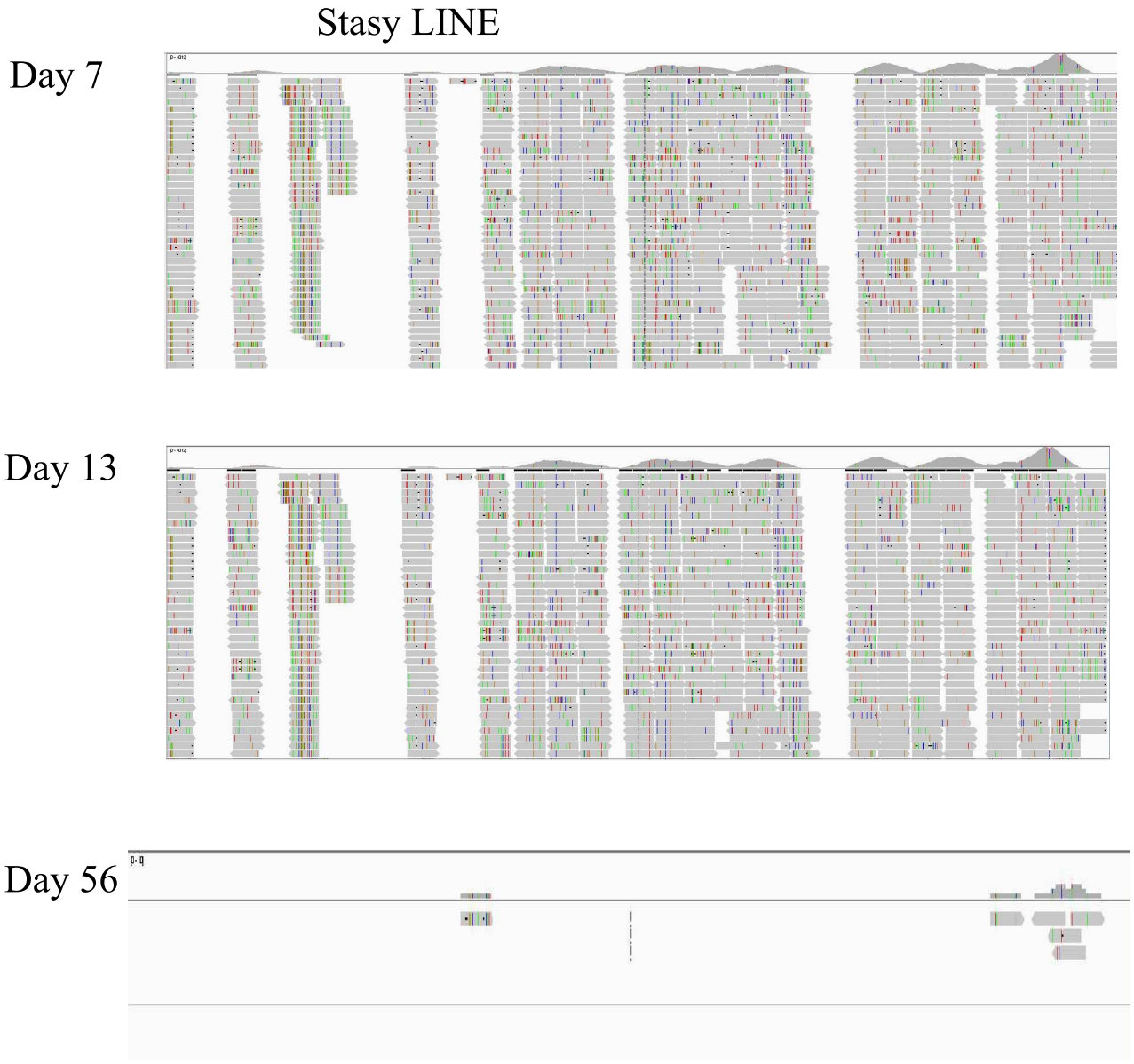


Figure 11. The pattern that the Illumina reads map to on the LINE Stasy after infection with *Mycosphaerella graminicola*. At day 7 and day 13 practically the entire element is expressed. The nature of the gaps is unclear and could be due to possibly poor quality or variability of the reference sequence.

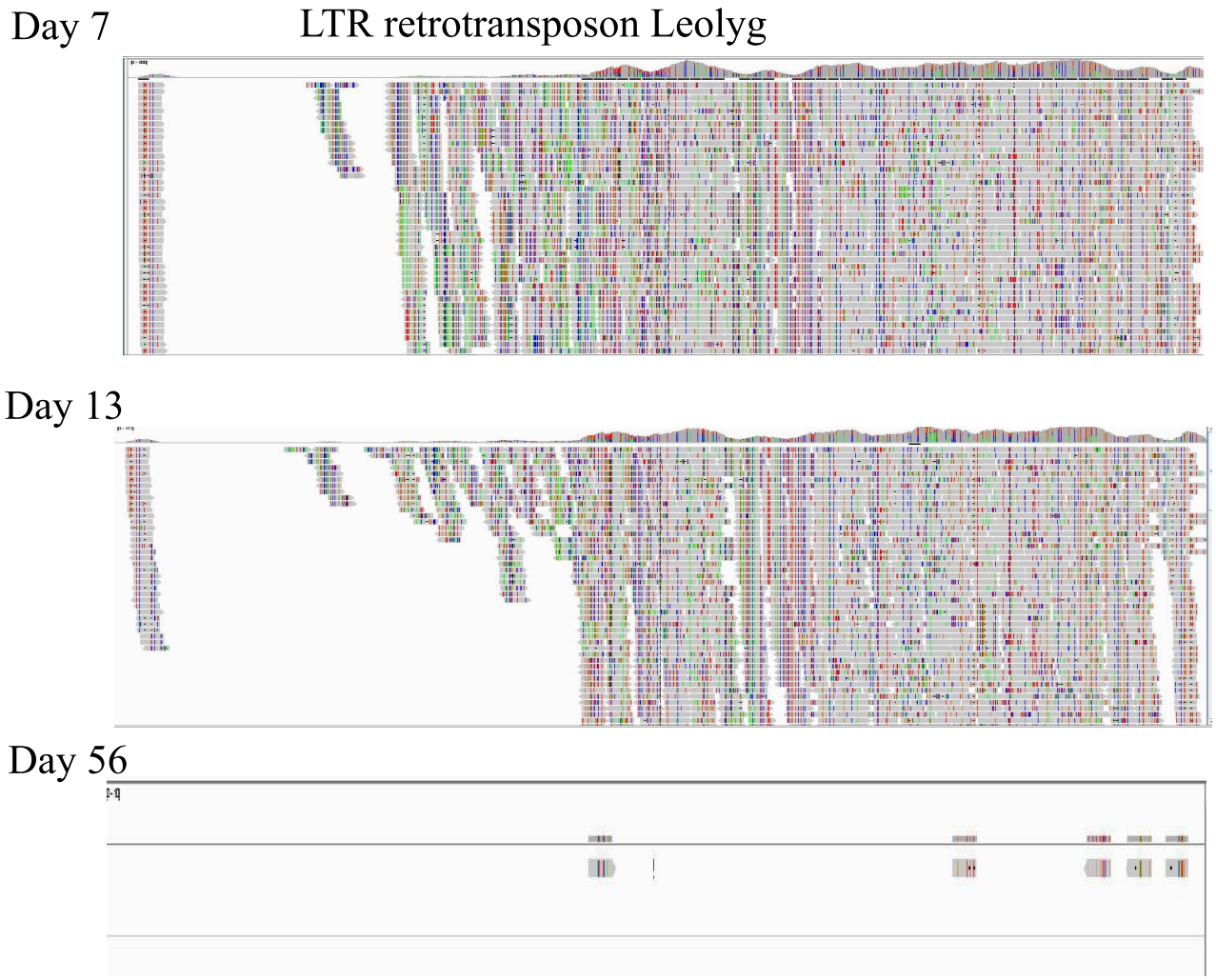


Figure 12. The pattern that the Illumina reads map to on the LTR retrotransposon Leolyg after infection with *Mycosphaerella graminicola*, with heavy expression seen in the region corresponding to the CDS.

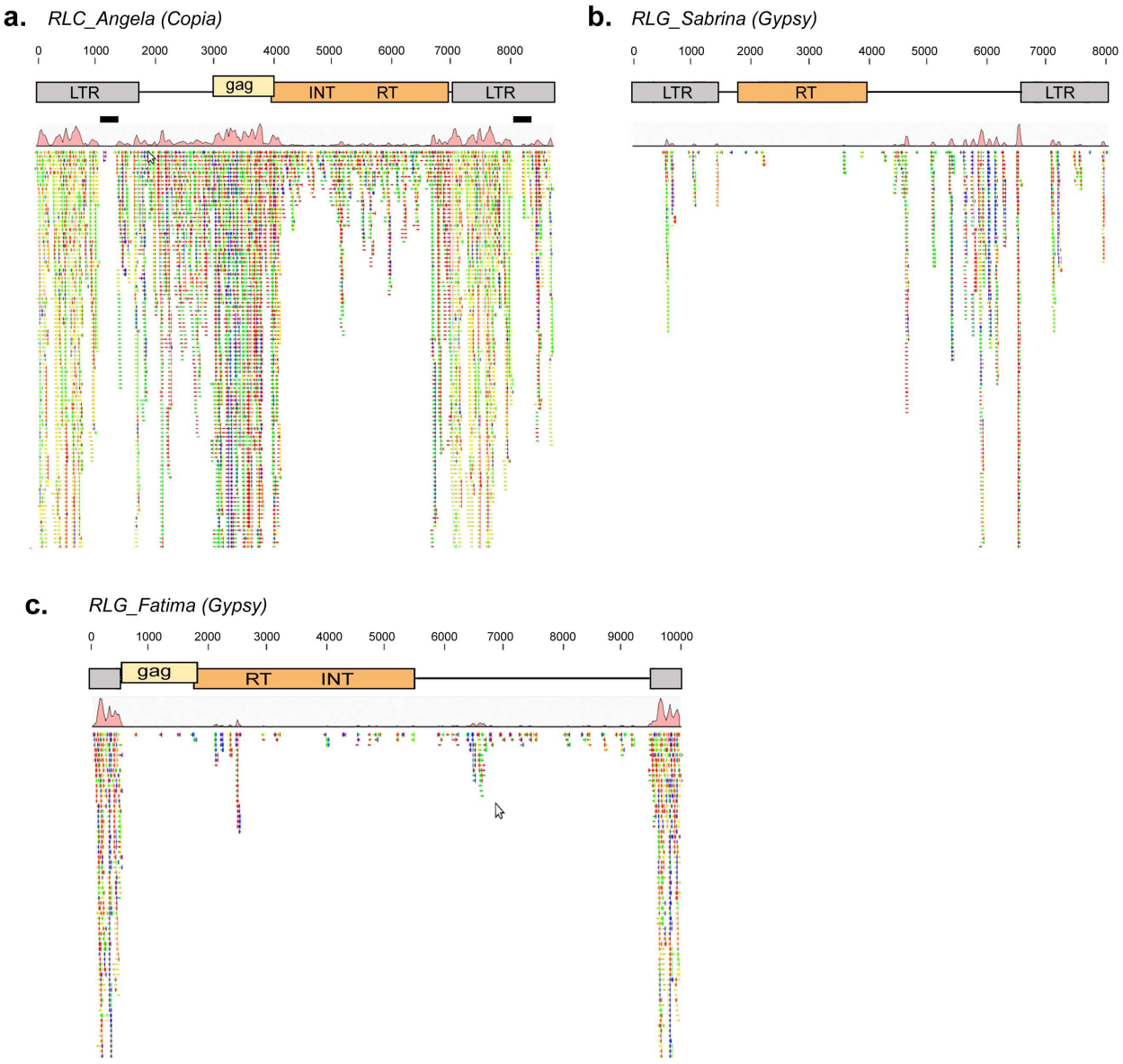


Figure 13. Expression patterns of the LTR retrotransposons *Angela*, *Sabrina* and *Fatima* in *T. aestivum* infected with *B. graminis*. a) *Angela* shows heavy expression in the CDS for gag and the LTRs, while the INT/RT region is less expressed. b) *Sabrina* shows lower expression in RT domain and the LTR than for the other region. c) *Fatima* shows high expression for the LTRs, but not the other regions of the element.

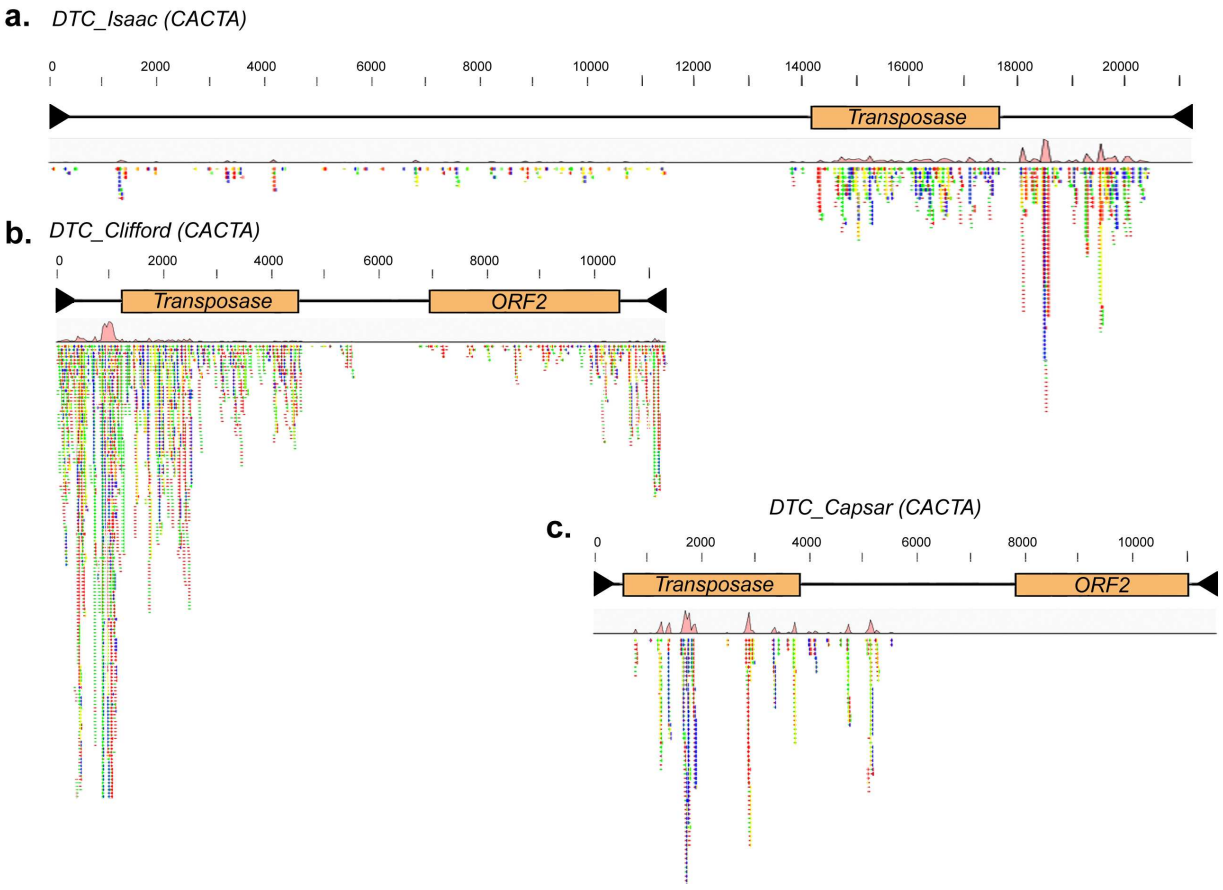


Figure 14. Expression patterns of the CACTA transposons *Caspar*, *Isaac* and *Clifford* in *T. aestivum* infected with *B. graminis*. a) *Isaac* shows expression of the transposase gene and the region downstream, suggesting there could be an additional ORF. b) *Clifford* and *Caspar* show strong expression of the transposase gene

5.4 Discussion

Expression patterns vary between different transposable elements of wheat after infection with pathogens. However, from the expression patterns presented here only limited conclusions could be drawn. Even though there are replicates for the RNA sequencing data produced when *T. aestivum* was infected with *Mycosphaerella graminicola*, there was no base line with RNA being extracted from uninfected wheat plants. This was the same situation for the RNA sequencing conducted on wheat infected with *Blumeria graminis*. Additionally, no replicates were taken for each of the time points. These limitations make it difficult to draw specific conclusions from the available data, but they still provide an idea of the how transposable elements are expressed or silenced under pathogen invasion. To our knowledge this is the first time a study has been conducted in Triticeae using RNA sequencing to investigate the influence that pathogens have in terms of changes undergone by transposable elements during infection.

Unexpected high expression of MITE elements

In both the RNA sequencing data obtained from *T. aestivum* infected with both *M. graminicola* and *B. graminis* it was found that the nonautonomous MITE *Thalos* showed the highest level of expression. This can be seen as unusual as MITEs are non-autonomous and do not encode a transcriptase. Expression would be expected in the CDS of autonomous elements, while the tandem inverted repeats (TIRs) and promoters are not expressed. As MITEs are essentially just TIRs it would be expected that no expression is seen. However, it was found the reads from the RNA sequencing data produced mapped to the entire length of the MITE. There are a few possible reasons for this: first, we can speculate that these elements are possibly found in a region that is highly expressed, for example in the downstream region of a gene. As MITEs are frequently found close to genes, this could be a likely explanation (Bureau *et al.*, 1994). Another possible reason is they could have their own internal promoter, that drives their expression. This has been described for ribosomal RNAs and these are transcribed by RNA polymerase III. There the promoter for 5S RNA transcription was found to occur between posi-

tion 55 and 80 within the gene and it was found that this small region can initiate transcription. However, it is difficult with the current data to suggest which of these possible explanations could be correct and further studies are needed to decipher the expression of MITEs.

Expression patterns of retroelements

Several LINEs were found to be highly expressed, although LINEs are found to be quite low in abundance in the genome of *T. aestivum* (Middleton *et al.*, 2013). This is in contrast to LTR retrotransposons, which form more than 50% of the overall genome content, with these being the least expressed in this small sample of elements investigated here. The expression patterns of LINEs were found to be consistent with expression of the whole element. In contrast, differences in expression patterns were found for the most abundant LTR retrotransposons. In the case of the LTR retrotransposon *Angela* expression was found to occur along the entire length of the element, with the LTRs and the gag ORF being the most heavily expressed. It is unknown why the LTR region (which is the promoter) would be so heavily expressed, but the gag gene is highly expressed as it is the structural protein that forms the coat of the virus-like particle that covers the reversetranscriptase/RNA complex. Therefore, much more of the gag protein is required than of reverse transcriptase and intergrase. By looking at the expression pattern of *Angela* elements it would appear that these elements are expressed.

In contrast to *Angelas* the LTR retrotransposon *Sabrina* shows a different pattern of expression, with the reverse transcriptase ORF showing practically no expression and distinct short segments showing high expression. This suggests that the small regions represent siRNAs and are involved in silencing, with such types of silencing, being previously described in *A. thaliana* (Qi *et al.*, 2006). The element *Fatima* also displays a different pattern of expression with LTRs being heavily expressed and the ORFs gag/RT/INT only in small regions, this suggests that the short segments in the CDS represent siRNAs. However, the reason for the high expression in the LTR region is unknown.

Expression patterns of DNA transposons

CACTA transposons show very diverse expression patterns and also differ in the overall amounts of reads that could be mapped to these elements. This was studied in *T. aestivum* infected with *B. graminis*, with the *CACTA* elements *Isaac*, *Caspar* and *Clifford* being investigated. These were found to have different expression profiles depending upon the element studied.

When looking at the overall pattern of expression between the elements and at the different time points of the infection in both infections with *M. graminicola* and *B. graminis*, it was found that the expression patterns remain consistent and the same regions of the element are expressed independently of the phase of the disease.

This limited sample of RNAseq data from TEs revealed a wide variety of expression patterns. Some patterns are consistent with the hypothesis of active or silenced elements, while others remain obscure (e.g the expression of MITEs and high expression of the LTR sequences). Furthermore, much broader studies will be required to explain the various patterns and to survey the whole scope of TE expression profiles.

6 General discussion

Next generation sequencing (NGS) was used in this study to investigate several aspects of the Triticeae genomes. Here we used NGS to analyse twelve species, with our main goals being to look at the genome composition of each of the species and in particular to focus on the variation of the different transposable element families within the genomes. A phylogeny of the species could also be derived from the assembly of the chloroplast genomes. Additionally, Triticeae divergence times could be estimated from this data. A further investigation was carried out to assess the dynamics of transposable elements, during pathogen infection, using transcriptome data. We found that NGS provides valuable data to investigate a multitude of topics in genome dynamics.

6.1 Transposable element composition in Triticeae

Full genome sequence information for the Triticeae are only just starting to emerge, with the draft sequences of *Triticum urartu* (Ling *et al.*, 2013), *Aegilops tauschii* (Jia *et al.*, 2013) and *Triticum aestivum* (Brenchley *et al.*, 2012) being released within the last year. However, due to the highly repetitive nature of these genomes, with approximately 80% of them being made up from transposable elements, it makes producing a completely assembled genome in Triticeae very challenging (Bennett *et al.*, 1976). A way to sample and assess the complexity of these genomes is to make use of smaller datasets using next generation sequencing. 454 pyrosequencing is a useful cost effective tool to study the genomes of Triticeae, as one run of 454 sequencing can provide sequence information of approximately 2.5% of the overall genome. Inferences can then be made into the composition of the genome as a whole, including identifying transposable elements, organelle sequences, genes, rDNAs and tRNAs. As transposable elements form the largest proportion of the genome, the 2% coverage provided by the 454 sequencing still gives a good indication of the overall TE composition. The use of the publically available *T. aestivum* (Brenchley *et al.*, 2012) 454 genome sequence supports this idea, as they performed a large number of runs of 454 sequencing. By taking

random sets of 454 sequences and testing them using BLAST searches, we were able to show that each run produced an almost identical result (our unpublished data).

The *BARE1* retrotransposon is found in the *Hordeum* species, whereas a closely related *Copia* element *Angela* is found in the genomes of wheat and its close relatives. Both *BARE1* and *Angela* elements were found in *S. cereale*. By drawing phylogenies based on the first 300 bp of the long tandem repeat (LTR) sequence it was possible to show that the *Angela* element is derived from the *BARE1* element. Thus our data could document the emergence of a new TE subfamily in a group of species.

The changes betweeen taxa in the repetitive element composition of the genome also occur rapidly, with some elements colonising the genome and increasing in their abundance rapidly, while others are reduced in number. Looking at these changes of the numbers and type of transposable elements within the genome, gives an indication of some of the processes that underlie how these changes in closely related taxa come about. The changes in transposable element composition could be due to several reasons, these include abiotic and biotic stress, as these have both been shown to influence the dynamics of transposable elements (Grandbastien, 1998). This can be seen in the A genome containing taxa *T. urartu*, *T. boeoticum* and *T. monococcum*, where the *Gypsy* element *Erika* is prevelant in relatively high abundance, with between 3 and 4% of the genomes of these taxa being formed by *Erika* elements. This is in contrast to the other Triticeae species in which typically approximately 1.5% or less of the genome is made up from *Erika* elements.

However, how these elements come to colonise the genomes of a species in a relatively short evolutionary time is still not understood. Transposons also have an influence on the size of genomes and it was found in a wild relative of cultivated rice, where the rice genome underwent recent bursts of transposable element activity resulting in a near doubling of the genome size in rice, without polyploidisation (Piegu *et al.*, 2006). There are many things that can influence the abundance of transposable elements in the genomes. RNA silencing will try to limit the expression of TEs in the genome, epigenetics also plays a role in silencing of TEs (Qi *et al.*, 2006). Increases in TE abundance could be due to the fact that Triticeae species undergo in-

tergressive events, such as gene flow, hybridisation and horizontal gene transfer between the closely related species, this could lead to some TEs becoming active in the new host and increasing their abundance before the host finds a means to silence them (Charles *et al.*, 2008).

6.2 Phylogeny of the Triticeae based on chloroplast sequences

By analysing a large segment of the chloroplast genomes it was possible to infer some information about the exact phylogeny of the Triticeae species. Several attempts have been made to correctly place the ancestral species of *T. aestivum* and multiple different positions of the A, B and D genomes in the tree have been suggested. Previous studies have focused on nuclear genes or just a small sample of genes from the chloroplast genome. Petersen *et al*,. 2006 used two nuclear genes DCM1 and EF-G from each genome of hexaploid wheat, from each of the A and B genomes of tetraploid *T. turgidum* and from each of the diploid ancestors. They also included the plastid gene *ndh*F and they found that the tree topology varied between the nuclear genes and the plastid gene. Their analysis of the plastid gene gave the same tree topology that was obtained from the large chloroplast sequence used in this study (B,(A,D). However, the variation was seen when the two nuclear genes produced another tree topology of the ancestral genomes, (D,(A,B). This is in contrast to the findings of Escobar *et al*,. 2011 in which they used 26 nuclear genes and one chloroplast gene and yet found another topology for *T. urartu*, *Ae. speltoides* and *Ae. tauschii* when compared to wheat.

A possible explanation for the differences seen in topology, could be due to reticulate evolution and by taking regions that are heavily influenced by introgression can have an impact on the phylogeny, by inferring these events rather than when the lineage actually split (Mason-Gamer, 2004). By using a large sequence from the chloroplast it is possible to remove the significance of introgression, as chloroplast genomes are clonal. Although the mutation rate is much lower than nuclear genes, they should still give a good indication to the phylogenic relationships between closely related species. However, using sequences from the chloroplast is only one way to look at Triticeae phylogeny. Gene sequences, rDNAs and other intergenic

sequences could also be used to obtain the phylogeny and with the sequence information of the Triticeae increasing, it should be possible to infer phylogeny using different haplotypes and cultivars. It is possible that the complete phylogeny of the Triticeae will be difficult to resolve, with the phylogeny being a mosaic rather than defined, with different loci or organelles having varying evolutionary histories.

6.3 Inheritance of chloroplast sequences in polyploid species

It has been suggested several times that the donor of the hexaploid *T. aestivum* chloroplast is either a closely related species to *Ae. speltoides* or a now extinct ancestor of this species. This was shown by comparing gene loci within the chloroplasts (Kilian *et al.*, 2007b). In our study we used a much larger 37 kb sequence of the large single copy region. This enabled us to draw a phylogenetic tree based on this large sequence to ascertain which of the three genome donors *T. urartu* (A), *Ae. speltoides* (B) and *Ae. tauschii* (D) donated its chloroplast genome to *T. aestivum*. By obtaining the phylogenetic tree, it was possible to provide evidence that the chloroplast of *T. aestivum* was probably donated by *Ae. speltoides* or a close now extinct ancestor.

Two further tetraploid taxa were included in this study, these were *Ae. cylindrica* (CD) and *Ae. geniculata* (MU). By analysing the chloroplast sequences it was possible to find the chloroplast donor of *Ae. cylindrica*, which is closely related to *Ae. tauschii*. Therefore, it is likely that the chloroplast genome of *Ae. cylindrica* was donated by *Ae. tauschii*.

The use of chloroplast sequences enable us to trace the lineage of the chloroplast to identify the possible chloroplast donors of polyploid Triticeae species. The same analysis could be done using mitochondrial genome sequences. Comparing the resulting trees could reveal incongruences between the mitochondrial and chloroplast evolutionary lineages.

6.4 Molecular dating

Using chloroplast genomes to obtain divergence dates can be problematic and several aspects have to be taken into consideration when calculating divergence times. One of the major considerations is that substitution rates vary markedly along the length of the chloroplast sequence, with there being differences in the nucleotide substitution rates between individual genes and the intergenic sequences contained within the chloroplast. However, no statistical difference was noted in nucleotide substitution rates between synonymous and nonsynonymous sites in the chloroplast sequence (Muse *et al.*, 1997; Duchene *et al.*, 2013b). This can lead to bias when employing dating techniques to the chloroplast genome sequences and care has to be taken to avoid this. Several methods used in molecular dating can be deployed to reduce the impact that different substitution rates between different sequences can have. One of these involves using a partioned model, where each gene can be assigned a different substitution model to account for the variation in substitution rates between different genes and intergenic sequences (Rutschmann, 2006). There are examples where whole chloroplast sequences have been used to date the divergence times of species, these include apple (Nikiforova *et al.*, 2013) and dicots and monocots (Chaw *et al.*, 2004). It was found that using chloroplast sequences was an effective method to draw phylogeny and infer divergence times. Therefore we believe that the use of chloroplast sequences to date the divergence of Triticeae is legitimate. By using penalised likelihood and Bayesian inference to date the divergence of the Triticeae, it was found that they produced almost identical results, the same was also found when the 37 kb sequence used was partioned and different substitution rates were applied to each gene or intergenic sequence in the partition. Therefore, the use of chloroplast sequences to date the divergence times of Triticeae is a valid method for dating the times of divergence in a closely related tribe such as the Triticeae. The main weak points for using sequences such as genes, chloroplasts and rDNAs to date the divergence of the Triticeae, is that there is only one calibration point and this is based upon the divergence of maize and rice, which was found to have occured 60 million years ago. It is from this point that all the divergence time estimates in the pooideae are estimated. This is possibly adequate for dating of the Triticeae

and gives consistent dates of divergence However, the introduction of new calibration points may shift the estimated absolute dates of divergence, while the relative divergence times will probably remain the same.

6.5 Transcriptome analysis of transposable elements

A limited dataset of RNA sequences from wheat infected with powdery mildew and septoria was used to analyse TE expression patterns and the possible nature of transposon silencing. Under each of the conditions and time points different TEs displayed different patterns of expression, as well as different parts of the elements being expressed. The most abundant element in the *T. aestivum* genome the *Copia* element *Angela* showed an expected pattern of expression, with the LTRs and ORFs being expressed, suggesting that this element is expressed during pathogen attack. Other elements investigated showed very different patterns of expression. The elements *Caspar* and *Sabrina* both show evidence of RNA silencing, as the reads produced only mapped to small regions of the element. However, much more extensive analysis is necessary to survey the whole range of TE expression patterns. Nevertheless our data provided a first insight into a surprising variety of expression patterns in TEs.

6.6 Future outlook

Here we showed that even with low coverage, 454 sequencing can be used in several ways to gain an understanding of how genomes are composed and how they evolve. Using next generation sequencing and transcriptome analysis enabled many new insights to be gained into the evolution and genome dynamics that underlie the tribe of the Triticeae. Due to its relatively low cost, a great number of species can be sampled, including sub-species and a number of different cultivars. Future studies should see the inclusion of a large number of additional species from this tribe, it would be interesting to see hows these are placed into the phylogeny and how close their dates of divergence are between each other. Further insights

could also be gained from the plethora of information generated by using next generation sequencing, such as fully assembled genomes of a multitude of species as the time and cost to produce complete sequences decreases. This would enable whole genome comparisons between closely related species and the identification of traits from wild species that could be then utilised in breeding programmes to improve crop species. However, there are current limitations with how much data can be stored and processed and there will be lots of bioinformatics challenges to deal with such a vast array of data.

References

AGI (2000) Analysis of the genome sequence of the flowering plant *Arabidopsis thaliana*. *Nature*, **408**, 796–815.

Akhunov, E.D., Akhunova, A.R. and Dvorak, J. (2007) Mechanisms and rates of birth and death of dispersed duplicated genes during the evolution of a multigene family in diploid and tetraploid wheats. *Mol Biol Evol*, **24**, 539–550.

Akhunov, E.D., Goodyear, A.W., Geng, S., Qi, L.L., Echalier, B., Gill, B.S., Miftahudin, Gustafson, J.P., Lazo, G., Chao, S. *et al.* (2003) The organization and rate of evolution of wheat genomes are correlated with recombination rates along chromosome arms. *Genome Res*, **13**, 753–763.

Arbogast, B.S., Edwards, S.V., Wakeley, J., Beerli, P. and Slowinski, J.B. (2002) Estimating divergence times from molecular data on phylogenetic and population genetic timescales. *Ann Rev Ecol Syst*, **33**, 707–740.

Baulcombe, D. (2004) RNA silencing in plants. *Nature*, **431**, 356–363.

Bennett, M.D. and Leitch, I.J. (2011) Nuclear DNA amounts in angiosperms: targets, trends and tomorrow. *Ann Bot*, **107**, 467–590.

Bennett, M.D. and Smith, J.B. (1976) Nuclear DNA amounts in angiosperms. *Philos Trans R Soc Lond B Biol Sci*, **274**, 227–274.

Bordbar, F., Rahiminejad, R.M., Saeida, H. and Blattner., F.R. (2011) Phylogeny and genetic diversity of D-genome species of *Aegilops* and *Triticum* (Triticeae, Poaceae) from Iran based on microsatellites, ITS and *trnL-F*. *Plant Syst Evol*, **291**, 117–131.

Bortiri, E., Coleman-Derr, D., Lazo, G.R., Anderson, O.D. and Gu, Y.Q. (2008) The complete chloroplast genome sequence of *Brachypodium distachyon*: sequence comparison and phylogenetic analysis of eight grass plastomes. *BMC Res Notes*, **1**, 61–69.

Bossolini, E., Wicker, T., Knobel, P.A. and Keller., B. (2007) Comparison of orthologous loci from small grass genome *Brachypodium* and rice: implications for wheat genomics and grass genome annotation. *Plant journal*, **49**, 704–717.

Brenchley, R., Spannagl, M., Pfeifer, M., Barker, G.L.A., D'Amore, R., Allen, A.M., McKenzie, N., Kramer, M., Kerhornou, A., Bolser, D. *et al.* (2012) Analysis of the bread wheat genome using whole-genome shotgun sequencing. *Nature*, **491**, 705–710.

Buckler, E.S., Thornsberry, J.M. and Kresovich, S. (2001) Molecular diversity, structure and domestication of grasses. *Genet Res*, **77**, 213–218.

Bureau, T.E. and Wessler, S.R. (1994) Stowaway: a new family of inverted repeat elements associated with the genes of both monocotyledonous and dicotyledonous plants. *Plant Cell*, **6**, 907–916.

Caldwell, K.S., Dvorak, J., Lagudah, E.S., Akhunov, E., Luo, M.C., Wolters, P. and Powell, W. (2004) Sequence polymorphism in polyploid wheat and their D-genome diploid ancestor. *Genetics*, **167**, 941–947.

Casacuberta, J.M. and Santiago, N. (2003) Plant LTR-retrotransposons and MITEs: control of transposition and impact on the evolution of plant genes and genomes. *Gene*, **311**, 1–11.

Cavalli-Sforza, L.L. (2005) The Human Genome Diversity Project: past, present and future. *Nat Rev Genet*, **6**, 333–340.

Chalupska, D., Lee, H.Y., Faris, J.D., Evrard, A., Chalhoub, B., Haselkorn, R. and Gornicki, P. (2008) Acc homoeoloci and the evolution of wheat genomes. *Proc Natl Acad Sci U S A*, **105**, 9691–9696.

Charles, M., Belcram, H., Just, J., Huneau, C., Viollet, A., Couloux, A., Segurens, B., Carter, M., Huteau, V., Coriton, O. *et al.* (2008) Dynamics and Differential Proliferation of Transposable Elements During the Evolution of the B and A Genomes of Wheat. *Genetics*, **180**, 1071–1086.

Chaw, S.M., Chang, C.C., Chen, H.L. and Li, W.H. (2004) Dating the monocot-dicot divergence and the origin of core eudicots using whole chloroplast genomes. *J Mol Evol*, **58**, 424–441.

Choulet, F., Wicker, T., Rustenholz, C., Paux, E., Salse, J., Leroy, P., Schlub, S., Paslier, M.C.L., Magdelenat, G., Gonthier, C. *et al.* (2010) Megabase level sequencing reveals contrasted organization and evolution patterns of the wheat gene and transposable element spaces. *Plant Cell*, **22**, 1686–1701.

Cohen, L. and Eyal, Z. (1993) The histology of processes associated with the infection of resistant and susceptible wheat cultivars with Septoria tritici. *Plant pathology*, **45**, 737–743.

Consortium, I.B.G.S., Mayer, K.F.X., Waugh, R., Brown, J.W.S., Schulman, A., Langridge, P., Platzer, M., Fincher, G.B., Muehlbauer, G.J., Sato, K. *et al.* (2012) A physical, genetic and functional sequence assembly of the barley genome. *Nature*, **491**, 711–716.

Dangl, J.L. and Jones, J.D. (2001) Plant pathogens and integrated defence responses to infection. *Nature*, **411**, 826–833.

Devos, K.M., Ma, J., Pontaroli, A.C., Pratt, L.H. and Bennetzen, J.L. (2005) Analysis and mapping of randomly chosen bacterial artificial chromosome clones from hexaploid bread wheat. *Proc Natl Acad Sci U S A*, **102**, 19243–19248.

Doebley, J.F., Gaut, B.S. and Smith, B.D. (2006) The molecular genetics of crop domestication. *Cell*, **127**, 1309–1321.

Drummond, A.J. and Rambaut, A. (2007) BEAST: Bayesian evolutionary analysis by sampling trees. *BMC Evol Biol*, **7**, 214.

Dubcovsky, J. and Dvorak, J. (2007) Genome plasticity a key factor in the success of polyploid wheat under domestication. *Science*, **316**, 1862–1866.

Duchene, D. and Bromham, L. (2013a) Rates of molecular evolution and diversification in plants: chloroplast substitution rates correlate with species-richness in the Proteaceae. *BMC evolutionary biology*, **13**, 65.

Duchene, D. and Bromham, L. (2013b) Rates of molecular evolution and diversification in plants: chloroplast substitution rates correlate with species-richness in the Proteaceae. *BMC Evol Biol*, **13**, 65.

Dvorak, J. and Akhunov, E. (2005) Tempos of gene locus deletions and duplications and their relationship to recombination rate during diploid and polyploid evolution in the *Aegilops-Triticum* alliance. *Genetics*, **171**, 323–332.

Eckardt, N.A. (2010) Evolution of domesticated bread wheat. *Plant Cell*, **22**, 993.

Edwards, D. and Batley, J. (2010) Plant genome sequencing: applications for crop improvement. *Plant Biotechnol J*, **8**, 2–9.

Edwards, S.V. and Beerli, P. (2000) Perspective: gene divergence, population divergence, and the variance in coalescence time in phylogeographic studies. *Evolution*, **54**, 1839–1854.

Eilam, T., Anikster, Y., Millet, E., Manisterski, J., Sagi-Assif, O. and Feldman, M. (2007) Genome size and genome evolution in diploid Triticeae species. *Genome*, **50**, 1029–1037.

Escobar, J.S., Scornavacca, C., Cenci, A., Guilhaumon, C., Santoni, S., Douzery, E.J.P., Ranwez, V., Glémin, S. and David, J. (2011) Multigenic phylogeny and analysis of tree incongruences in Triticeae (Poaceae). *BMC Evol Biol*, **11**, 181.

Feldman, M., Lupton, F. and Miller, T. (1995) *Evolution of crop Plants*. Longman Scientific, London (UK), 2^{nd} edition.

Feschotte, C., Jiang, N. and Wessler, S.R. (2002) Plant transposable elements: where genetics meets genomics. *Nat Rev Genet*, **3**, 329–341.

Gaut, B. (2002) Evolutionary dynamics of grass genomes. *New Phytol*, **154**, 15–28.

Ghaffary, S.M.T., Faris, J.D., Friesen, T.L., Visser, R.G.F., van der Lee, T.A.J., Robert, O. and Kema, G.H.J. (2012) New broad-spectrum resistance to septoria tritici blotch derived from synthetic hexaploid wheat. *Theor Appl Genet*, **124**, 125–142.

Giovanni, M.D., Cenci, A., Janni, M. and D'Ovidio, R. (2008) A LTR *Copia* retrotransposon and *Mutator* transposons interrupt *Pgip* genes in cultivated and wild wheats. *Theor Appl Genet*, **116**, 859–867.

Golovnina, K.A., Glushkov, S.A., Blinov, A.G., Mayorov, V.I., Adkison, L.R. and Goncharov, N.P. (2007) Molecular phylogeny of the genus *Triticum L*. *pl. Sys. Evol*, **264**, 195–216.

Goodwin, S.B. (2007) Back to basics and beyond: increasing the level of resistance to Septoria tritici blotch in wheat. *Australasian plant pathology*, **36**, 532–538.

Goodwin, S.B., M'barek, S.B., Dhillon, B., Wittenberg, A.H.J., Crane, C.F., Hane, J.K., Foster, A.J., der Lee, T.A.J.V., Grimwood, J., Aerts, A. *et al.* (2011) Finished genome of the fungal wheat pathogen *Mycosphaerella graminicola* reveals dispensome structure, chromosome plasticity, and stealth pathogenesis. *PLoS Genet*, **7**, e1002070.

Gouy, M., Guindon, S. and Gascuel, O. (2010) SeaView version 4: A multiplatform graphical user interface for sequence alignment and phylogenetic tree building. *Mol Biol Evol*, **27**, 221–224.

Grandbastien, M.A. (1998) Activation of plant retrotransposons under stress conditions. *Trends in plant science*, **3**, 181–187.

Haudry, A., Cenci, A., Guilhaumon, C., Paux, E., Poirier, S., Santoni, S., David, J. and Glémin, S. (2008) Mating system and recombination affect molecular evolution in four Triticeae species. *Genet Res (Camb)*, **90**, 97–109.

Hirosawa, S., Takumi, S., Ishii, T., Kawahara, T., Nakamura, C. and Mori, N. (2004) Chloroplast and nuclear DNA variation in common wheat: insight into the origin and evolution of common wheat. *Genes Genet Syst*, **79**, 271–282.

Hollister, J.D. and Gaut, B.S. (2007) Population and evolutionary dynamics of Helitron transposable elements in *Arabidopsis thaliana*. *Mol Biol Evol*, **24**, 2515–2524.

Huang, S., Sirikhachornkit, A., Faris, J.D., Su, X., Gill, B.S., Haselkorn, R. and Gornicki, P. (2002a) Phylogenetic analysis of the acetyl-CoA carboxylase and 3-phosphoglycerate kinase loci in wheat and other grasses. *Plant Mol Biol*, **48**, 805–820.

Huang, S., Sirikhachornkit, A., Su, X., Faris, J., Gill, B., Haselkorn, R. and Gornicki, P. (2002b) Genes encoding plastid acetyl-CoA carboxylase and 3-phosphoglycerate kinase of the *Triticum/Aegilops* complex and the evolutionary history of polyploid wheat. *Proc Natl Acad Sci U S A*, **99**, 8133–8138.

Initiative, I.B. (2010) Genome sequencing and analysis of the model grass *Brachypodium distachyon*. *Nature*, **463**, 763–768.

Isidore, E., Scherrer, B., Chalhoub, B., Feuillet, C. and Keller, B. (2005) Anciient haplotypes resulting from extensive molecular rearrangements in the wheat A genome have been maintained in species of three different ploidy levels. *Genome Res*, **15**, 526–536.

Jaenisch, R. and Bird, A. (2003) Epigenetic regulation of gene expression: how the genome integrates intrinsic and environmental signals. *Nat Genet*, **33 Suppl**, 245–254.

Jansen, R., Kaittanis, C., Saski, C., Lee, S.B., Tomkins, J., Alverson, A. and Daniell, H. (2006) Phylogenetic analyses of *Vitis* (Vitaceae) based on complete chloroplast genome sequences: effects of taxon sampling and phylogenetic methods on resolving relationships among rosids. *BMC Evol Biol*, **6**, 32.

Jia, J., Zhao, S., Kong, X., Li, Y., Zhao, G., He, W., Appels, R., Pfeifer, M., Tao, Y., Zhang, X. *et al.* (2013) *Aegilops tauschii* draft genome sequence reveals a gene repertoire for wheat adaptation. *Nature*, **496**, 91–95.

Johnson, B.L. and Dhailwal, H.S. (1976) Reproductive isolation of *Triticum boeoticum* and *Triticum urartu* and the origin of the Tetraploid Wheats. *American Journal of Botany*, **63**, 1088–1094.

Kalendar, R., Tanskanen, J., Immonen, S., Nevo, E. and Schulman, A.H. (2000) Genome

evolution of wild barley (*Hordeum spontaneum*) by *BARE-1* retrotransposon dynamics in response to sharp microclimatic divergence. *Proc Natl Acad Sci U S A*, **97**, 6603–6607.

Kapitonov, V.V. and Jurka, J. (2001) Rolling-circle transposons in eukaryotes. *Proc Natl Acad Sci U S A*, **98**, 8714–8719.

Kapitonov, V.V. and Jurka, J. (2007) Helitrons on a roll: eukaryotic rolling-circle transposons. *Trends Genet*, **23**, 521–529.

Kidwell, M.G. (2002) Transposable elements and the evolution of genome size in eukaryotes. *Genetica*, **115**, 49–63.

Kihara, H. (1944) Discovery of the DD analyser, one of the ancestors of *vulgare* wheats. *Agr & Hort,*, **19**, 889–890.

Kilian, B., Mammen, K., Millet, E., Sharma, R., Graner, A., Salamini, F., Hammer, K. and Özkan, H. (2011) Wild Crop Relatives: Genomic and Breeding Resources Cereals. *Aegilops* L. In: Kole C (editor). *Springer*, 1–76.

Kilian, B., Ozkan, H., Walther, A., Kohl, J., Dagan, T., Salamini, F. and Martin, W. (2007a) Molecular diversity at 18 loci in 321 wild and 92 domesticate lines reveal no reduction of nucleotide diversity during *Triticum monococcum* (Einkorn) domestication: implications for the origin of agriculture. *Mol Biol Evol*, **24**, 2657–2668.

Kilian, B., Özkan, H., Deusch, O., Effgen, S., Brandolini, A., Kohl, J., Martin, W. and Salamini, F. (2007b) Independent wheat B and G genome origins in outcrossing *Aegilops* progenitor haplotypes. *Mol Biol Evol*, **24**, 217–227.

Kilian, B., Özkan, H., Pozzi, C. and Salamini, F. (2009) *Genetics and Genomics of the Triticeae. Plant Genetics and Genomics: Crops and Models 7*. Springer Science+Business Media, LLC, New York.

Kilian, B., Özkan, H., Kohl, J., von Haeseler, A., Barale, F., Deusch, O., Brandolini, A., Yucel, C., Martin, W. and Salamini, F. (2006) Haplotype structure at seven barley genes:

relevance to gene pool bottlenecks, phylogeny of ear type and site of barley domestication. *Mol Genet Genomics*, **276**, 230–241.

Kimber, G. (1974) A reassessment of the origin of the polyploid wheats. *Genetics*, **78**, 487–492.

Kimber, G.K., S.E. (1987) *Evolution of the genus Triticeae and origin of cultivated wheat.* American Society of Agronomy, Inc., Crop Science Society of America, Inc., Soil Science Society of America, Inc., Madison, Wisconsin, USA.

Librado, P. and Rozas, J. (2009) DnaSP v5: a software for comprehensive analysis of DNA polymorphism data. *Bioinformatics*, **25**, 1451–1452.

Ling, H.Q., Zhao, S., Liu, D., Wang, J., Sun, H., Zhang, C., Fan, H., Li, D., Dong, L., Tao, Y. *et al.* (2013) Draft genome of the wheat A-genome progenitor *Triticum urartu*. *Nature*, **496**, 87–90.

Lloyd, A.H. and Timmis, J.N. (2011) The origin and characterization of new nuclear genes originating from a cytoplasmic organellar genome. *Mol Biol Evol*, **28**, 2019–2028.

Macas, J., Neumann, P. and Navratilova, A. (2007) Repetitive DNA in the pea (*Pisum sativum L*.) genome: comprehensive characterization using 454 sequencing and comparison to soybean and *Medicago truncatula*. *BMC Genomics*, **8**, 427.

Mar, J.C., Harlow, T.J. and Ragan, M.A. (2005) Bayesian and maximum likelihood phylogenetic analyses of protein sequence data under relative branch-length differences and model violation. *BMC Evol Biol*, **5**, 8.

Mardis, E.R. (2008) The impact of next-generation sequencing technology on genetics. *Trends Genet*, **24**, 133–141.

Mason-Gamer, R.J. (2004) Reticulate evolution, introgression, and intertribal gene capture in an allohexaploid grass. *Syst Biol*, **53**, 25–37.

Matsuo, M., Ito, Y., Yamauchi, R. and Obokata, J. (2005) The rice nuclear genome continuously integrates, shuffles, and eliminates the chloroplast genome to cause chloroplast-nuclear DNA flux. *Plant Cell*, **17**, 665–675.

McFadden, E. and Sears, E. (1946) The origin of *Triticum spelta* and its free-threshing hexaploid relatives. *J Hered*, **37**, 81–89.

Middleton, C.P., Stein, N., Keller, B., Kilian, B. and Wicker, T. (2013) Comparative analysis of genome composition in Triticeae reveals strong variation in transposable element dynamics and nucleotide diversity. *Plant J*, **73**, 347–356.

Mori, N., Liu, Y.G. and Tsunewaki, K. (1995) Wheat phylogeny determined by RFLP analysis of nuclear DNA. 2. Wild tetraploid wheats. *Theor Appl Genet*, **90**, 129–134.

Muse, S.V. and Gaut, B.S. (1997) Comparing patterns of nucleotide substitution rates among chloroplast loci using the relative ratio test. *Genetics*, **146**, 393–399.

Nei, M. and Li, W.H. (1979) Mathematical model for studying genetic variation in terms of restriction endonucleases. *Proc Natl Acad Sci U S A*, **76**, 5269–5273.

Nikiforova, S.V., Cavalieri, D., Velasco, R. and Goremykin, V. (2013) Phylogenetic analysis of 47 chloroplast genomes clarifies the contribution of wild species to the domesticated apple maternal line. *Mol Biol Evol*, **30**, 1751–1760.

Nock, C.J., Waters, D.L.E., Edwards, M.A., Bowen, S.G., Rice, N., Cordeiro, G.M. and Henry, R.J. (2010) Chloroplast genome sequences from total DNA for plant identification. *Plant Biotechnol J*.

Ogihara, Y., Isono, K., Kojima, T., Endo, A., Hanaoka, M., Shiina, T., Terachi, T., Utsugi, S., Murata, M., Mori, N. *et al.* (2002) Structural features of a wheat plastome as revealed by complete sequencing of chloroplast DNA. *Mol Genet Genomics*, **266**, 740–746.

Onodera, Y., Haag, J.R., Ream, T., Nunes, P.C., Pontes, O. and Pikaard, C.S. (2005) Plant nuclear RNA polymerase IV mediates siRNA and DNA methylation-dependent heterochromatin formation. *Cell*, **120**, 613–622.

Özkan, H., Tuna, M., Kilian, B., Mori, N. and Ohta, S. (2010) Genome size variation in diploid and tetraploid wild wheats. *AoB Plants*, **plq015**.

Paterson, A.H., Bowers, J.E., Bruggmann, R., Dubchak, I., Grimwood, J., Gundlach, H., Haberer, G., Hellsten, U., Mitros, T., Poliakov, A. *et al.* (2009a) The *Sorghum bicolor* genome and the diversification of grasses. *Nature*, **457**, 551–556.

Paterson, A.H., Bowers, J.E., Feltus, F.A., Tang, H., Lin, L. and Wang, X. (2009b) Comparative genomics of grasses promises a bountiful harvest. *Plant Physiol*, **149**, 125–131.

Peng, J., Richards, D.E., Hartley, N.M., Murphy, G.P., Devos, K.M., Flintham, J.E., Beales, J., Fish, L.J., Worland, A.J., Pelica, F. *et al.* (1999) 'Green revolution' genes encode mutant gibberellin response modulators. *Nature*, **400**, 256–261.

Petersen, G., Seberg, O., Yde, M. and Berthelsen, K. (2006) Phylogenetic relationships of *Triticum* and *Aegilops* and evidence for the origin of the A, B, and D genomes of common wheat (*Triticum aestivum*). *Mol Phylogenet Evol*, **39**, 70–82.

Piegu, B., Guyot, R., Picault, N., Roulin, A., Sanyal, A., Saniyal, A., Kim, H., Collura, K., Brar, D.S., Jackson, S. *et al.* (2006) Doubling genome size without polyploidization: dynamics of retrotransposition-driven genomic expansions in *Oryza australiensis*, a wild relative of rice. *Genome Res*, **16**, 1262–1269.

Pollak, E. (1987) On the theory of partially inbreeding finite populations. I. Partial selfing. *Genetics*, **117**, 353–360.

Posada, D. (2008) jModelTest: phylogenetic model averaging. *Mol Biol Evol*, **25**, 1253–1256.

Provan, J., Wolters, P., Caldwell, K.H. and Powell, W. (2004) High-resolution organellar genome analysis of *Triticum* and *Aegilops* sheds new light on cytoplasm evolution in wheat. *Theor Appl Genet*, **108**, 1182–1190.

Qi, Y., Hel, X., Wang, X.J., Kohany, O., Jurka, J. and Hannon, G.J. (2006) Distinct catalytic and non-catalytic roles of ARGONAUTE4 in RNA-directed DNA methylation. *Nature*, **443**, 1008–1012.

Rebollo, R., Horard, B., Hubert, B. and Vieira, C. (2010) Jumping genes and epigenetics: Towards new species. *Gene*, **454**, 1–7.

Rees, H. and Walters., M.R. (1965) Nuclear DNA and the evolution of wheat. *Heredity*, **20**, 73–82.

Rutschmann, F. (2006) Molecular dating of phylogenetic trees:A brief review of current methods that estimate divergence times. *Diversity Distrib J.*, **12**, 35–48.

Salamini, F., Ozkan, H., Brandolini, A., Schäfer-Pregl, R. and Martin, W. (2002) Genetics and geography of wild cereal domestication in the near east. *Nat Rev Genet*, **3**, 429–441.

Sanderson, M.J. (2002) Estimating absolute rates of molecular evolution and divergence times: a penalized likelihood approach. *Mol Biol Evol*, **19**, 101–109.

Sanderson, M.J. (2003) r8s: inferring absolute rates of molecular evolution and divergence times in the absence of a molecular clock. *Bioinformatics*, **19**, 301–302.

Saski, C., Lee, S.B., Fjellheim, S., Guda, C., Jansen, R.K., Luo, H., Tomkins, J., Rognli, O.A., Daniell, H. and Clarke, J.L. (2007) Complete chloroplast genome sequences of *Hordeum vulgare*, *Sorghum bicolor* and *Agrostis stolonifera*, and comparative analyses with other grass genomes. *Theor Appl Genet*, **115**, 571–590.

Schatz, M.C. (2012) Computational thinking in the era of big data biology. *Genome Biol*, **13**, 177.

Scherrer, B., Isidore, E., Klein, P., soon Kim, J., Bellec, A., Chalhoub, B., Keller, B. and Feuillet, C. (2005) Large intraspecific haplotype variability at the *Rph7* locus results from rapid and recent divergence in the barley genome. *Plant Cell*, **17**, 361–374.

Schnable, P.S., Ware, D., Fulton, R.S., Stein, J.C., Wei, F., Pasternak, S., Liang, C., Zhang, J., Fulton, L., Graves, T.A. *et al.* (2009) The B73 maize genome: complexity, diversity, and dynamics. *Science*, **326**, 1112–1115.

Senerchia, N., Wicker, T., Felber, F. and Parisod, C. (2013) Evolutionary dynamics of retrotransposons assessed by high-throughput sequencing in wild relatives of wheat. *Genome Biol Evol*, **5**, 1010–1020.

Shendure, J. and Ji, H. (2008) Next-generation DNA sequencing. *Nature Biotechnology*, **26**, 1135–1145.

Sheppard, A.E. and Timmis, J.N. (2009) Instability of plastid DNA in the nuclear genome. *PLoS Genet*, **5**, e1000323.

Slotkin, R.K. (2010) The epigenetic control of the *Athila* family of retrotransposons in *Arabidopsis*. *Epigenetics*, **5**, 483–490.

Slotkin, R.K. and Martienssen, R. (2007) Transposable elements and the epigenetic regulation of the genome. *Nat Rev Genet*, **8**, 272–285.

Soleimani, V.D., Baum, B.R. and Johnson, D.A. (2006) Quantification of the retrotransposon *BARE-1* reveals the dynamic nature of the barley genome. *Genome*, **49**, 389–396.

Soltis, D.E., Albert, V.A., Savolainen, V., Hilu, K., Qiu, Y.L., Chase, M.W., Farris, J.S., Stefanovic, S., Rice, D.W., Palmer, J.D. *et al.* (2004) Genome-scale data, angiosperm relationships, and "ending incongruence": a cautionary tale in phylogenetics. *Trends Plant Sci*, **9**, 477–483.

Steiper, M.E. and Young, N.M. (2008) Timing primate evolution lessons from the discordance between molecular and paleontological estimates. *Evolutionary anthropology*, **17**, 179–188.

Swaminathan, K., Varala, K. and Hudson, M. (2007) Global repeat discovery and estimation of genomic copy number in a large, complex genome using a high-throughput 454 sequence survey. *BMC Genomics*, **8**, 132.

Tajima, F. (1989) Statistical method for testing the neutral mutation hypothesis by DNA polymorphism. *Genetics*, **123**, 585–595.

Tamura, K., Peterson, D., Peterson, N., Stecher, G., Nei, M. and Kumar, S. (2011) MEGA5: Molecular Evolutionary Genetics Analysis using maximum likelihood, evolutionary distance, and maximum parsimony methods. *Mol Biol Evol*.

Tenaillon, M.I., Hufford, M.B., Gaut, B.S. and Ross-Ibarra, J. (2011) Genome size and transposable element content as determined by high-throughput sequencing in maize and *Zea luxurians*. *Genome Biol Evol*, **3**, 219–229.

Todorovska, E. (2007) Retrotransposons and their role in plant-genome evolution. *Biotechnology & Biotechnological Eq*, **21:3**, 294–305.

Tomita, M., Asao, M. and Kuraki, A. (2010) Effective isolation of retrotransposons and repetitive DNA families from the wheat genome. *J Integr Plant Biol*, **52**, 679–691.

Tsukahara, S., Kobayashi, A., Kawabe, A., Mathieu, O., Miura, A. and Kakutani, T. (2009) Bursts of retrotransposition reproduced in *Arabidopsis*. *Nature*, **461**, 423–426.

Tsunewaki, K. (1996) Plasmon analysis as the counterpart of genome analysis. *Methods of genome analysis in plants*, 271–299.

Tsunewaki, K. and Ogihara, Y. (1983) The molecular basis of genetic diversity among cytoplasms of *Triticum* and *Aegilops* species. II. on the origin of polyploid wheat cytoplasms as suggested by chloroplast DNA restriction fragment patterns. *Genetics*, **104**, 155–171.

Tsunewaki, K., Wang, G.Z. and Matsuoka, Y. (2002) Plasmon analysis of Triticum (wheat) and Aegilops. 2. Characterization and classification of 47 plasmons based on their effects on common wheat phenotype. *Genes Genet Syst*, **77**, 409–427.

Vicient, C.M., Suoniemi, A., Anamthawat-Jonsson, K., Tanskanen, J., Beharav, A., Nevo, E. and Schulman, A.H. (1999) Retrotransposon *BARE-1* and Its role in genome evolution in the genus *Hordeum*. *Plant Cell*, **11**, 1769–1784.

Vitte, C. and Panaud, O. (2005) LTR retrotransposons and flowering plant genome size: emergence of the increase/decrease model. *Cytogenet Genome Res*, **110**, 91–107.

Wang, X., Shi, X. and Hao, B. (2002) The transfer RNA genes in Oryza sativa L. ssp. indica. *Sci China C Life Sci*, **45**, 504–511.

Wicker, T., Krattinger, S.G., Lagudah, E.S., Komatsuda, T., Pourkheirandish, M., Matsumoto, T., Cloutier, S., Reiser, L., Kanamori, H., Sato, K. *et al.* (2009a) Analysis of intraspecies diversity in wheat and barley genomes identifies breakpoints of ancient haplotypes and provides insight into the structure of diploid and hexaploid triticeae gene pools. *Plant Physiol*, **149**, 258–270.

Wicker, T., Taudien, S., Houben, A., Keller, B., Graner, A., Platzer, M. and Stein, N. (2009b) A whole-genome snapshot of 454 sequences exposes the composition of the barley genome and provides evidence for parallel evolution of genome size in wheat and barley. *Plant J*, **59**, 712–722.

Wicker, T., Yahiaoui, N. and Keller, B. (2007) Illegitimate recombination is a major evolutionary mechanism for initiating size variation in plant resistance genes. *Plant J*, **51**, 631–641.

Yang, G., Nagel, D.H., Feschotte, C., Hancock, N. and Wessler, S. (2009) Tuned for transposition: Molecular determinants underlying the hyperactivity of a *Stowaway MITE.* Science*, **325**, 1391–1394.*

Yu, J., Hu, S., Wang, J., Wong, G.K.S., Li, S., Liu, B., Deng, Y., Dai, L., Zhou, Y., Zhang, X. et al. *(2002) A draft sequence of the rice genome (Oryza sativa L. ssp. indica).* Science*, **296**, 79–92.*

Özkan, H., Brandolini, A., Pozzi, C., Effgen, S., Wunder, J. and Salamini, F. *(2005) A reconsideration of the domestication geography of tetraploid wheats.* Theor Appl Genet*, **110**, 1052–1060.*

7 Acknowledgements

I would like to take my time to thank the many people who have helped me along on this sometimes arduous, but mainly fun journey of discovery. The last four years, producing my PhD thesis have been an interesting time, where I've learnt many things both professionally and personally. I would like to extend my graditude to all the people that have made this possible.
Firstly I would like to thank PD Dr. Thomas Wicker for his superb supervision and advice and Prof. Dr. Beat Keller for allowing me the oppertunity to do my PhD thesis within his group and to Prof. Dr Robert Dudler for being a member of my commitee. Benjamin Kilian and Nils Stein for collaborating, providing data and extensive comments on the manuscripts.
Special thanks also have to go to the members, both past and present of P3-12, my time here has been made far more enjoyable, even if its just working in the office or enjoying a beer on the roof.
Thomas Wicker, Stefan Roffler, Margarita Shatalina, Kostas Kristas, Simone Oberhansli, Fabrizio Menardo, Jan Buchmann and James Breen. I would like to thank all of you. I couldn't ask for better lab mates.

I would like to thank all the members of the P3 lab, for all your help and support. It has been great to work in such a friendly and open lab enviroment. Its been great working with all of you.

My gratitude also has to extend to my friends overseas. To Stewart Giddins, David Hughes, Cory Blose, Micheal Lawrence, Daniel Ackerley and Tina-Louise Carroll, for the meals, nights out, laughs and random camping adventures and to all my friends in the UK.
My family have been really important in getting my into this position. I would like to take this time to thank them all for their help and support. My Granparents Brian and Freda, my parents Richard and Karen and my sister Louise and my brother James.
I would like to this oppertunity to thank Dominique Braun for her love and support over the last few months and for providing me with enough cake to feed an army. I really appreciate everything you've done for me, I know it hasn't been easy, but I am eternally grateful.
I guess there isn't much more I can say except, "so long and thanks for all the fish".